데이터와 분포 #2

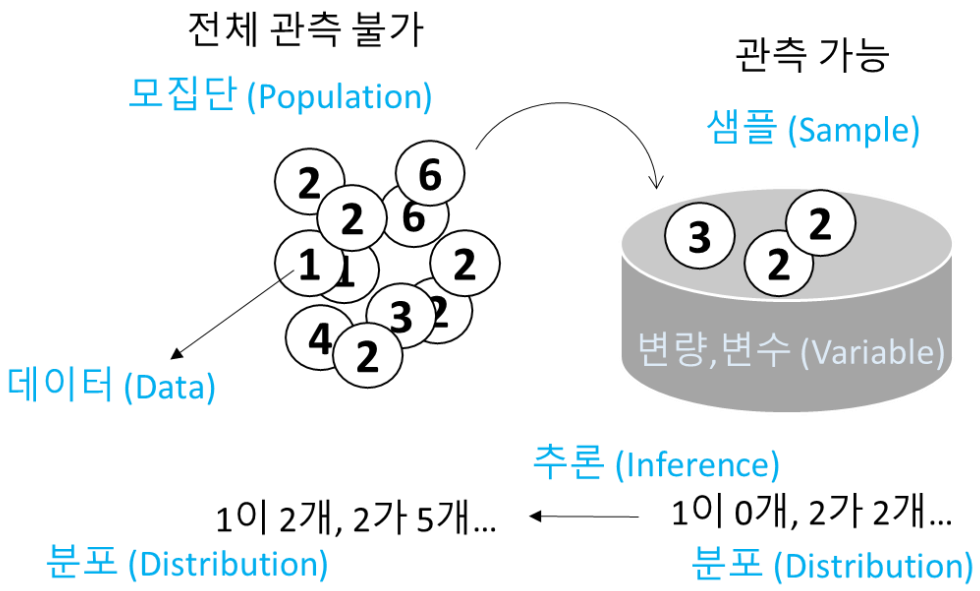
과학기술연합대학원대학교
한국생명공학연구원 스쿨
시스템생명공학전공

haseong@kribb.re.kr
김하성

https://github.com/greendaygh/2019-USTOT-stat

Summary of lecture #1

통계
데이터, 정보
일변량
요약통계량
-Center: mean, median..
-Spread: variance, range..
-Shape: skewness, ..



Bivariate data

1. Independent data (unpaired data) - Similarity
2. Paired data - Relationship
3. Bivariate categorical data

} Numerical

1. Independent data (Unpaired data)

독립적인 두 변수의 데이터 비교

- 학교간 성적 비교
- 나라별 GDP 비교
- 5살, 6살 아이들 키 비교
- 유산균 효과 비교

Independent samples

- Common experimental setup – a cohort dataset
  - Treatment (case) vs. Control
  - Placebo effect in control group

  ex) What food can have an impact on sports performance?

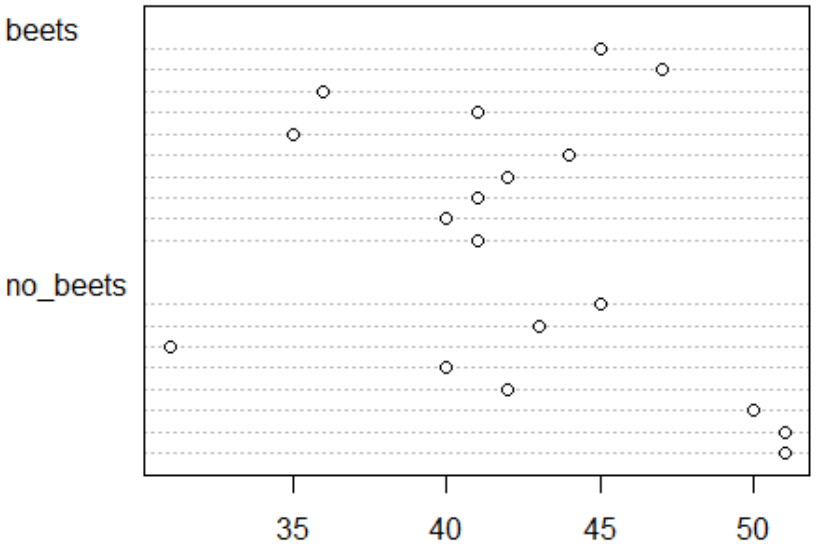
beets: 41, 40, 41, 42, 44, 35, 41, 36, 47, 45
no_beets: 51, 51, 50, 42, 40, 31, 43, 45

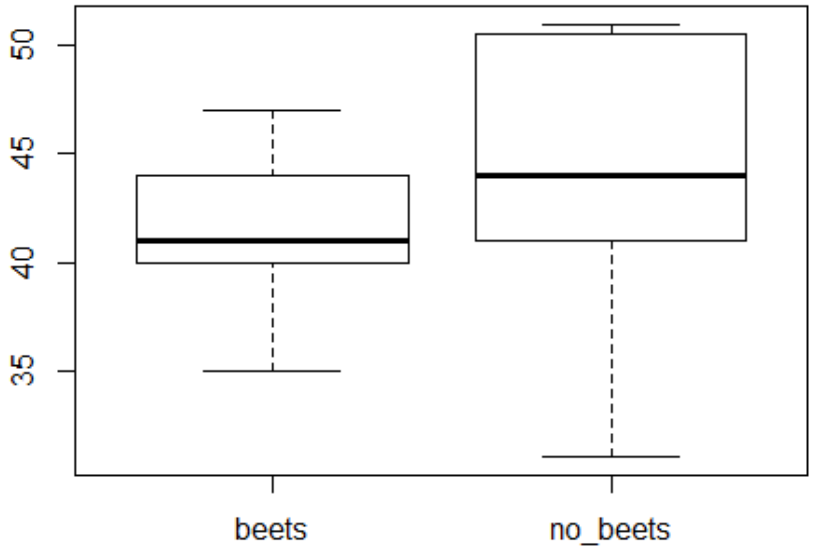
- Three longest and one shortest times in no beets-eaten group
- Similar population?
- Similar center?
- Similar spread?
- Same shape?

Plots for comparison

beets: 41, 40, 41, 42, 44, 35, 41, 36, 47, 45
no_beets: 51, 51, 50, 42, 40, 31, 43, 45

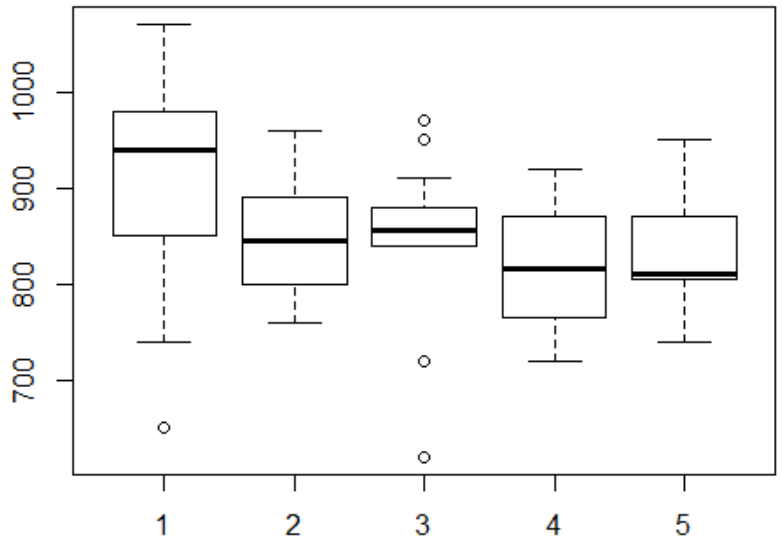
xbar_beets: 41.2, sd_beets: 3.70
xbar_no_beets: 44.12, sd_no_beets: 6.81

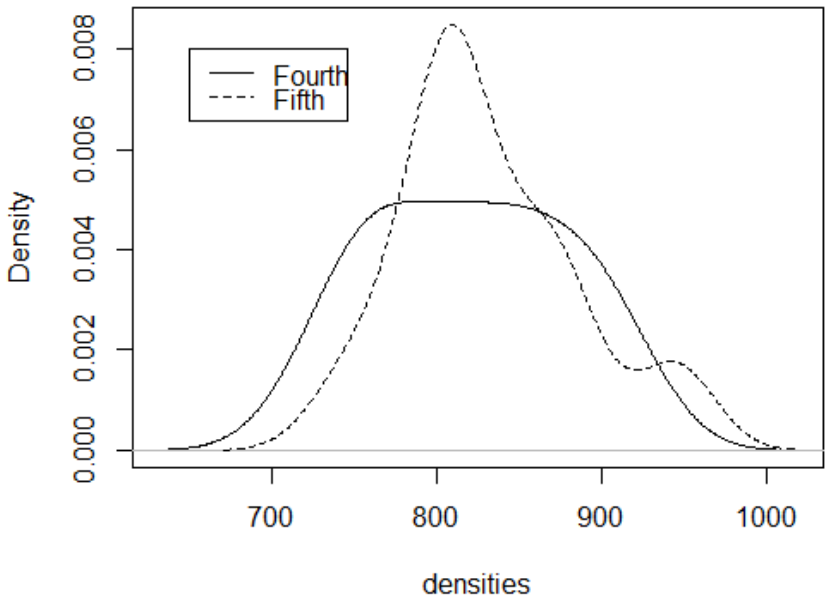




Density plot (shape)

Ex) Data set of the speed of light in air (Michelson in 1879). The data set consists of five experiments

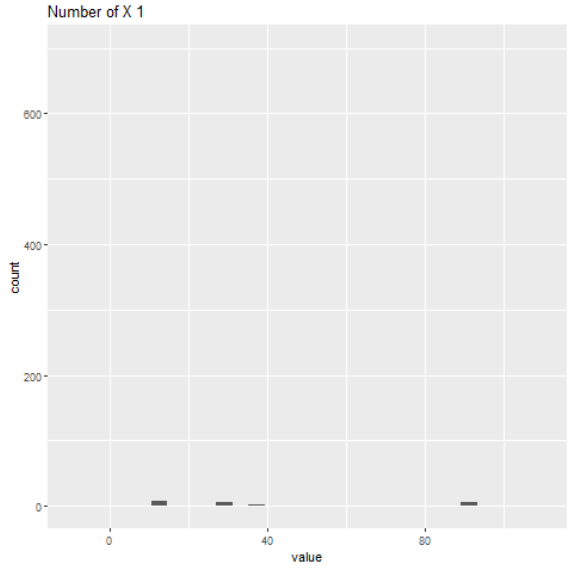


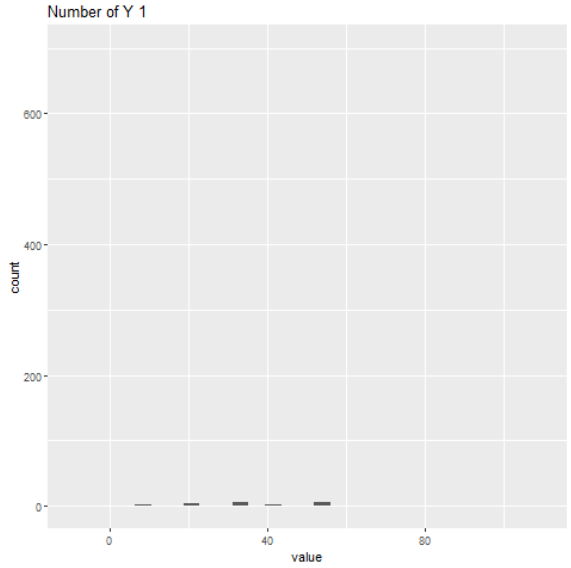


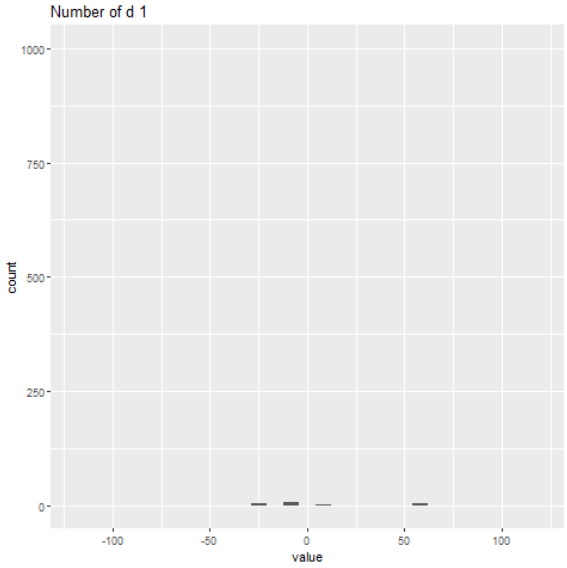
Difference of center

$$d = \bar{x} - \bar{y}$$

$$T = \frac{d - E(d|H_0)}{SE(d|H_0)}$$





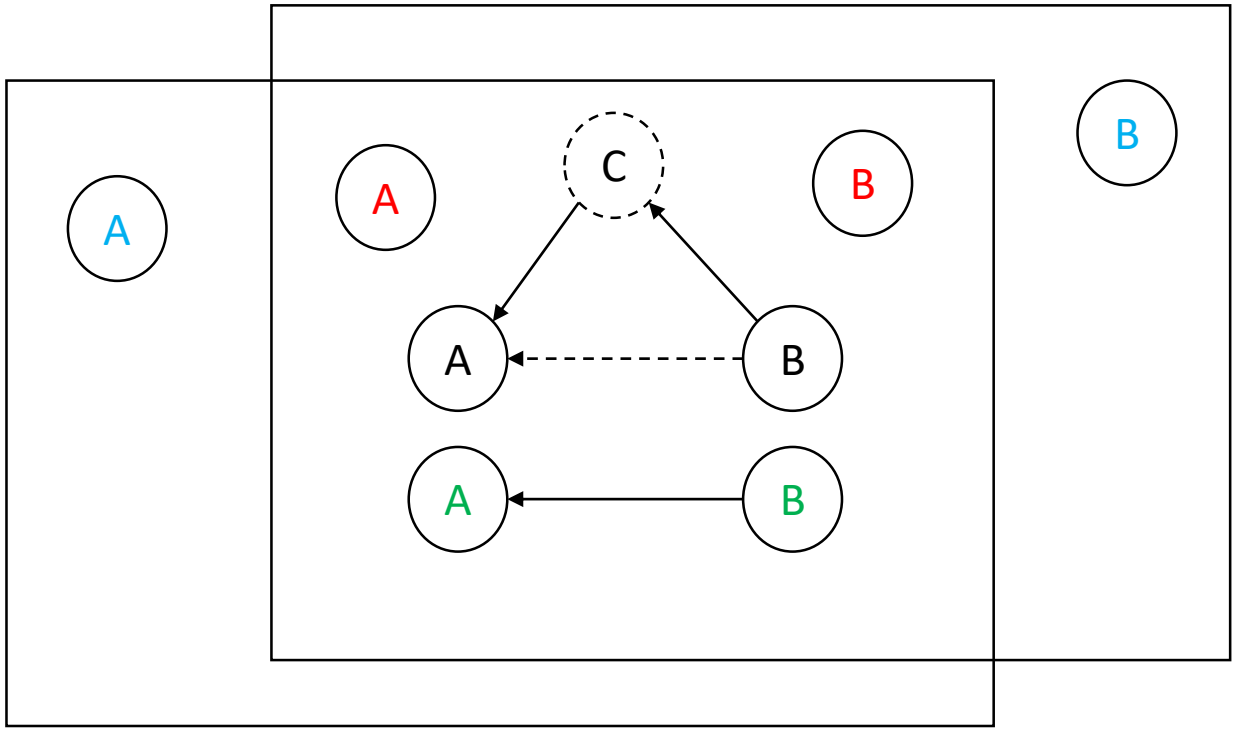


2. Paired data

같은 샘플에서 두 변수 측정 비교

- 수학을 잘하는 학생들이 과학도 잘한다
- 키가 큰 아이들이 몸무게도 많이 나간다

두 변수간 관계 (relationship)



Independence (독립)
Correlation (상관)
Association (연관)
Causation (인과)

Gulliver's Travels (1726)

Part I: A Voyage to Lilliput

그리고 내 오른쪽 엄지 둘레도 재었는데 그게 끝이었다. 이들은 수학적인 계산을 통해, 엄지 둘레를 두 배로 곱해서 손목 둘레를 알아내고, 이런 식으로 목 둘레와 허리 둘레까지 계산해 냈다. 거기에 본을 뜨라고 입고 있던 셔츠를 바닥에 펼쳐 주었더니 재봉사들은 내 치수를 정확히 알아내었다. 그리고 본격적으로 옷을 만드는 데는 재단사 300명이 동원되었다.

걸리버 여행기Bestseller World's Classics 1

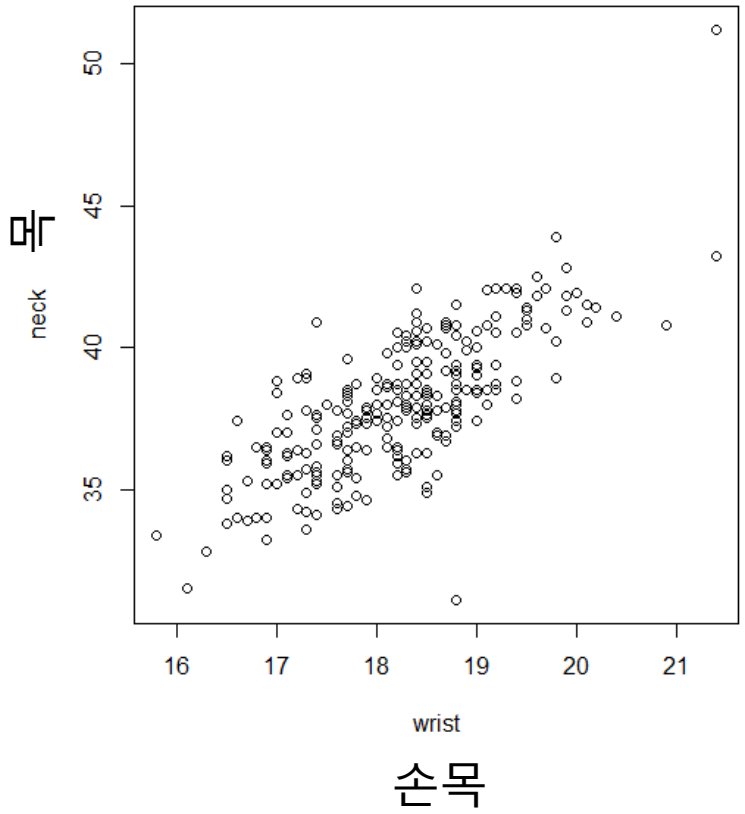
Paired data

- Two different variables from one sample
  e.g) height and weight of a student
  vs) heights and weights of students
- fat dataset
  - Different body dimensions for a cohort of 252 males
  - Aiming to predict body fat index

$$(x_1, y_1), (x_2, y_2), \dots, (x_{252}, y_{252})$$

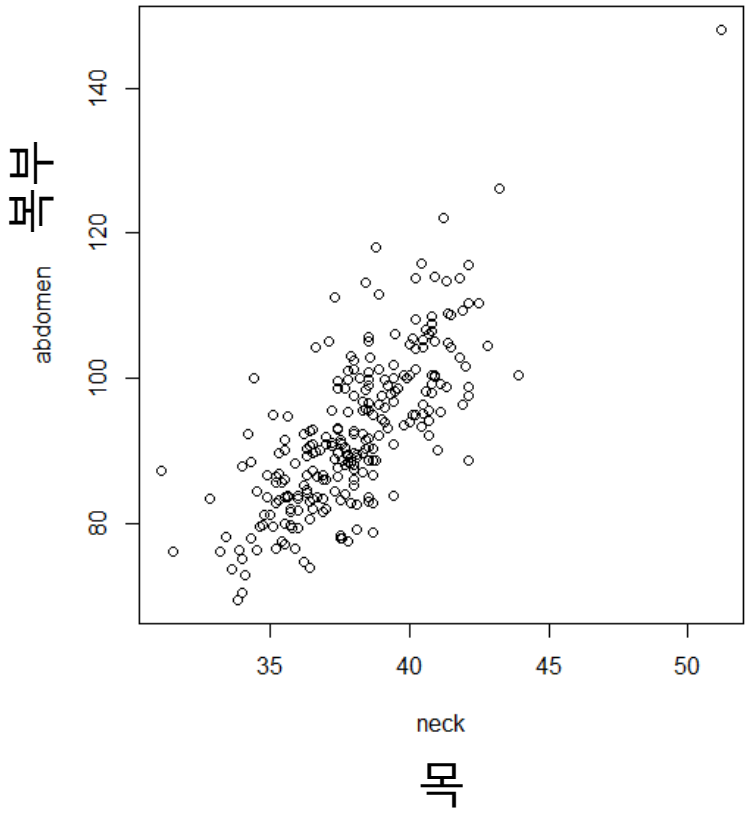
	case	body.fat	body.fat.siri	density	age	weight	height	BMI	ffweight	neck	chest	abdomen	hip	thigh	knee	ankle	bicep	forearm	wrist
1	1	12.6	12.3	1.0708	23	154.25	67.75	23.7	134.9	36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	17.1
2	2	6.9	6.1	1.0853	22	173.25	72.25	23.4	161.3	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2
3	3	24.6	25.3	1.0414	22	154.00	66.25	24.7	116.0	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6
4	4	10.9	10.4	1.0751	26	184.75	72.25	24.9	164.7	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
5	5	27.8	28.7	1.0340	24	184.25	71.25	25.6	133.1	34.4	97.3	100.0	101.9	63.2	42.2	24.0	32.2	27.7	17.7
6	6	20.6	20.9	1.0502	24	210.25	74.75	26.5	167.0	39.0	104.5	94.4	107.8	66.0	42.0	25.6	35.7	30.6	18.8
7	7	19.0	19.2	1.0549	26	181.00	69.75	26.2	146.6	36.4	105.1	90.7	100.3	58.4	38.3	22.9	31.9	27.8	17.7
8	8	12.8	12.4	1.0704	25	176.00	72.50	23.6	153.6	37.8	99.6	88.5	97.1	60.0	39.4	23.2	30.5	29.0	18.8
9	9	5.1	4.1	1.0900	25	191.00	74.00	24.6	181.3	38.1	100.9	82.5	99.9	62.9	38.3	23.8	35.9	31.1	18.2
10	10	12.0	11.7	1.0722	23	198.25	73.50	25.8	174.4	42.1	99.6	88.6	104.1	63.1	41.7	25.0	35.6	30.0	19.2
11	11	7.5	7.1	1.0830	26	186.25	74.50	23.6	172.3	38.5	101.5	83.6	98.2	59.7	39.7	25.2	32.8	29.4	18.5
12	12	8.5	7.8	1.0812	27	216.00	76.00	26.3	197.7	39.4	103.6	90.9	107.7	66.2	39.2	25.9	37.2	30.2	19.0
13	13	20.5	20.8	1.0513	32	180.50	69.50	26.3	143.5	38.4	102.0	91.6	103.9	63.4	38.3	21.5	32.5	28.6	17.7
14	14	20.8	21.2	1.0505	30	205.25	71.25	28.5	162.5	39.4	104.1	101.8	108.6	66.0	41.5	23.7	36.9	31.6	18.8
15	15	21.7	22.1	1.0484	35	187.75	69.50	27.4	147.0	40.5	101.3	96.4	100.1	69.0	39.0	23.1	36.1	30.5	18.2
16	16	20.5	20.9	1.0512	35	162.75	66.00	26.3	129.3	36.4	99.1	92.8	99.2	63.1	38.7	21.7	31.1	26.4	16.9
17	17	28.1	29.0	1.0333	34	195.75	71.00	27.3	140.8	38.9	101.9	96.4	105.2	64.8	40.8	23.1	36.2	30.8	17.3

Lilliputians' hypothesis



목

손목



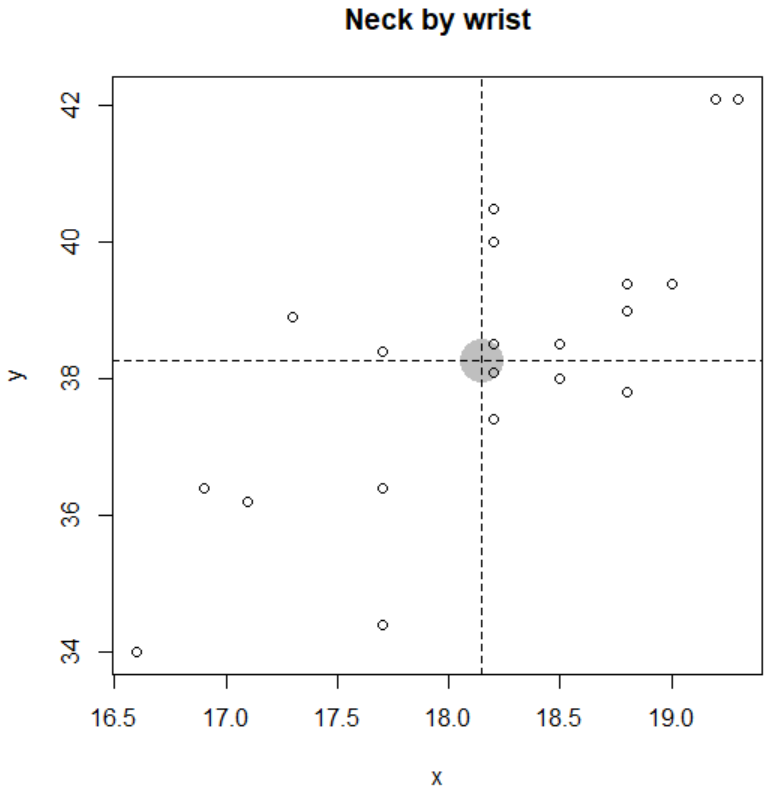
복부

목

* Scatter plot (산점도)

Correlation (상관성)

- Numeric summary for how closely related are the two variables
- The strength of a linear relationship
- For the definition, shift the center to mean



가운데 점을 기준으로 4등분
상관성 높은 데이터일수록
1,3 또는 2,4 사분면에 위치

Covariance (공분산)

- 서로 다른 변수들이 서로 얼마나 의존하는지에 대한 수치적 표현
- 공분산은 X의 편차와 Y의 편차의 곱의 평균
- 선형관계 관련성 측정

$$cov(x, y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

- Cov(X, Y) > 0 X가 증가 할 때 Y도 증가한다.(양의 상관관계)
- Cov(X, Y) < 0 X가 증가 할 때 Y는 감소한다.(음의 상관관계)
- X, Y 선형관계 없을 경우 Cov(X, Y) = 0 , 그러나 Cov(X, Y)=0 라 해서 선형관계 없는 것은 아님

공분산의 단점 - 값의 범위가 정해져 있지 않음

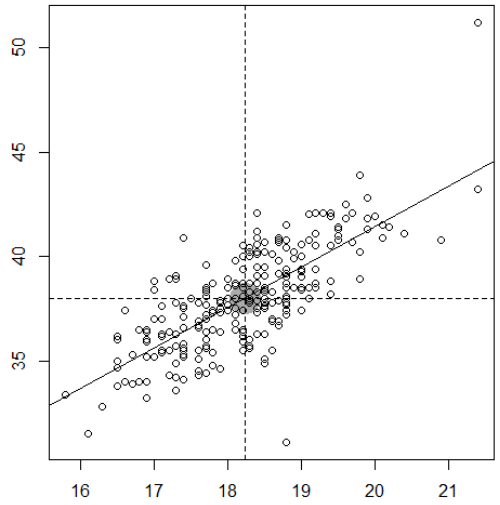
Pearson correlation coefficient (상관계수)

Pearson's r

- X의 z 값과 Y의 z값의 곱의 평균
- 선형관계 관련성 측정

$$cor(x,y) = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

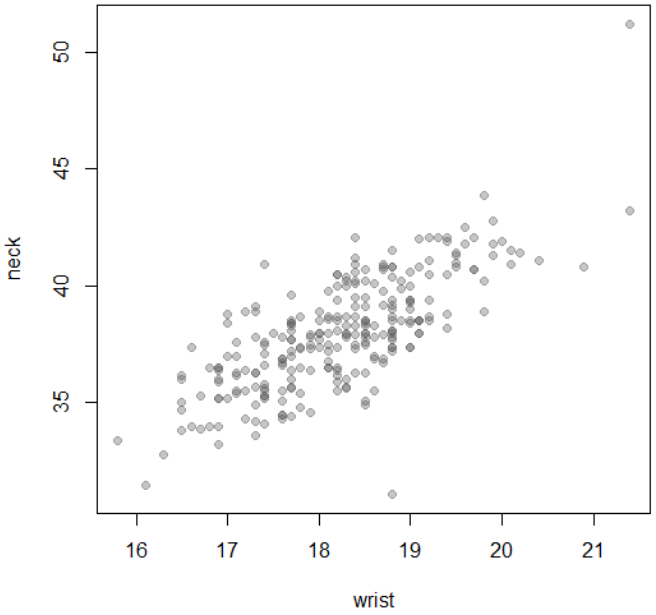
Y

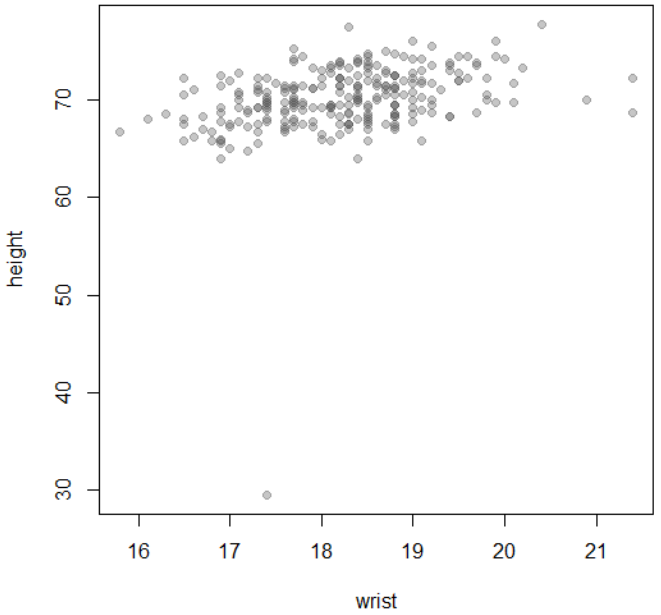


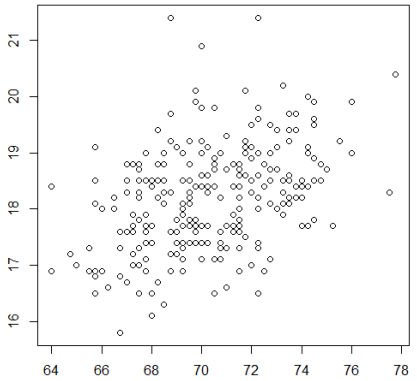
X

-1 ≤ r ≤ 1
-1: 강한 음의 선형 상관
1: 강한 양의 선형 상관

ex) 손목둘레 vs. 목둘레, 손목둘레 vs. 키







Cov(X, Y) = 1.69

Cor(X, Y) = 0.74

Cov(X, Y) = 1.10

Cov(X, Y)' = 0.97

Cor(X, Y) = 0.32

Cor(X, Y)' = 0.39

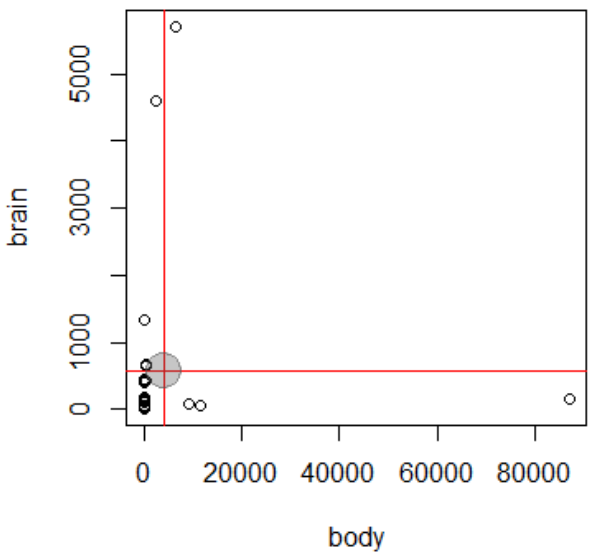
Why?

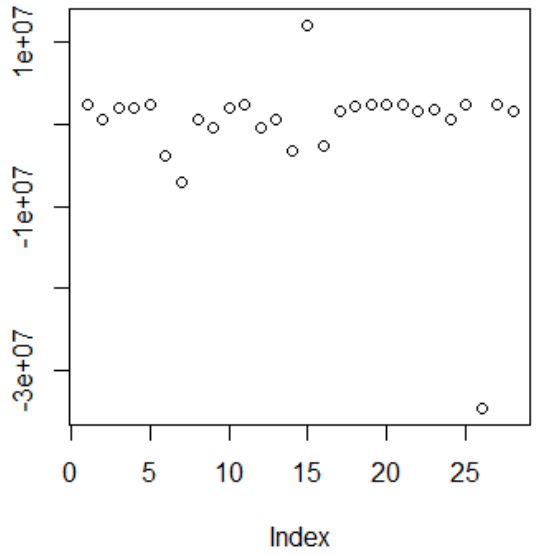
Body weight vs. Brain size

	body	brain
Mountain beaver	1.350	8.1
Cow	465.000	423.0
Grey wolf	36.330	119.5
Goat	27.660	115.0
Guinea pig	1.040	5.5
Dipliodocus	11700.000	50.0
Asian elephant	2547.000	4603.0
Donkey	187.100	419.0
Horse	521.000	655.0
Potar monkey	10.000	115.0
Cat	3.300	25.6
Giraffe	529.000	680.0
Gorilla	207.000	406.0
Human	62.000	1320.0
African elephant	6654.000	5712.0
Triceratops	9400.000	70.0
Rhesus monkey	6.800	179.0
Kangaroo	35.000	56.0
Golden hamster	0.120	1.0
Mouse	0.023	0.4
Rabbit	2.500	12.1
Sheep	55.500	175.0
Jaguar	100.000	157.0
Chimpanzee	52.160	440.0
Rat	0.280	1.9
Brachiosaurus	87000.000	154.5
Mole	0.122	3.0
Pig	192.000	180.0

Cor(Body, Brain) = -0.005341

$(x_i - \bar{x})(y_i - \bar{y})$





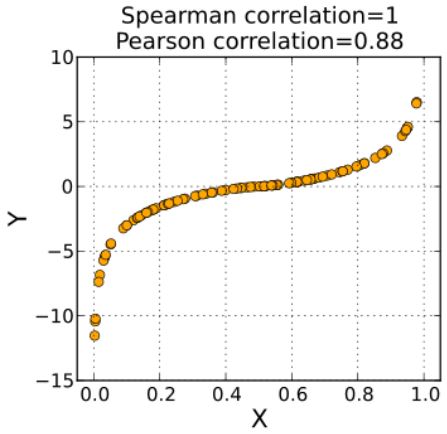
한 데이터 포인트의 큰 cross product 값이
다른 데이터의 값들을 (정보를) 제거

(Spearman's rank) Correlation

- Monotonically related data will be 1 or -1
- Applicable both of linear and nonlinear data
- Rank the data and take the correlation

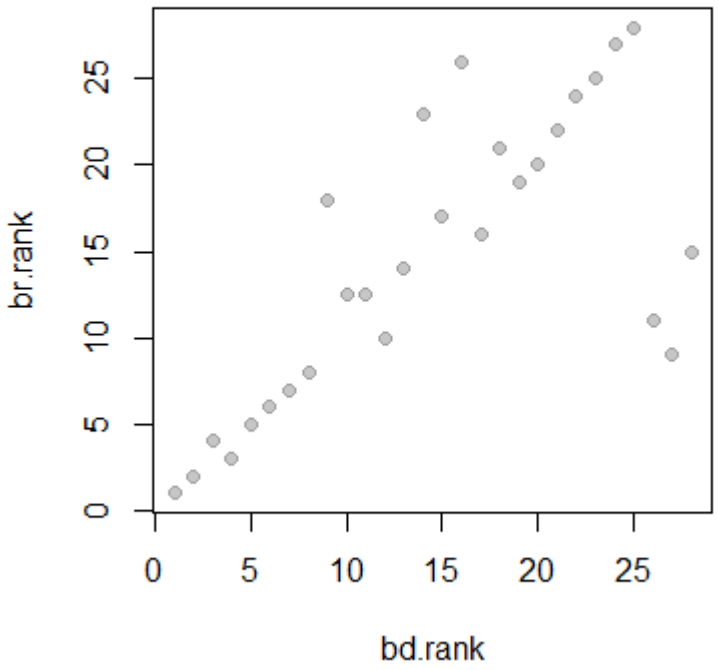
$$cov(x,y) = \frac{1}{n-1}\sum\left(R(x_i) - \overline{R(x)}\right)\left(R(y_i) - \overline{R(y)}\right)$$

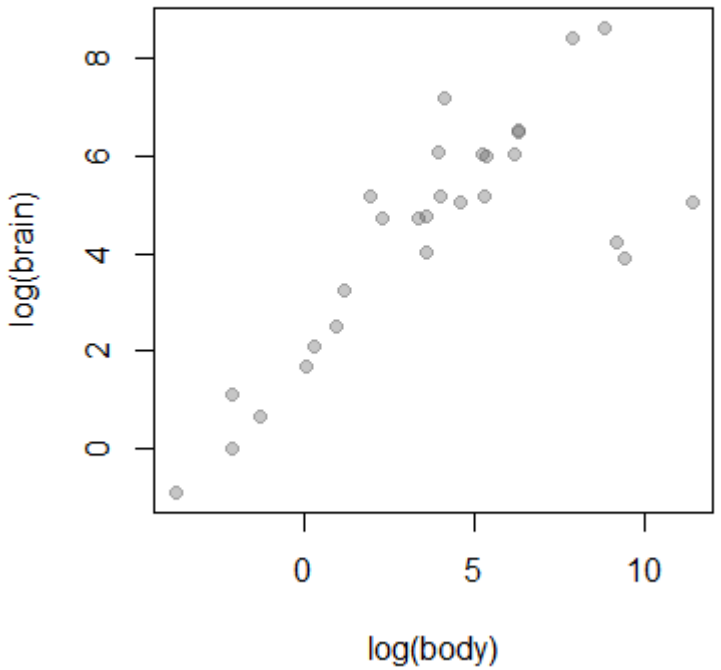
$$cor(x,y) = \frac{1}{n-1}\sum\left(\frac{R(x_i) - \overline{R(x)}}{s_{R(x)}}\right)\left(\frac{R(y_i) - \overline{R(y)}}{s_{R(y)}}\right)$$



-wiki-

Body weight vs. Brain size



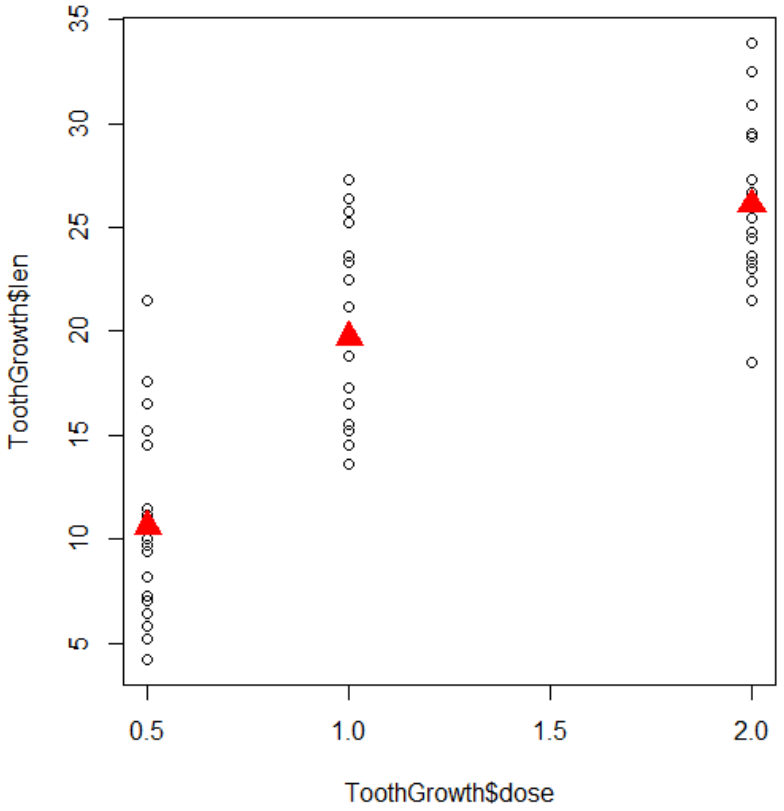


cor(rank(body), rank(brain)) = 0.7162994

cor(log(Body), log(Brain)) = 0.7794935

Correlated averages with replication

- correlation of averaged data vs. whole data



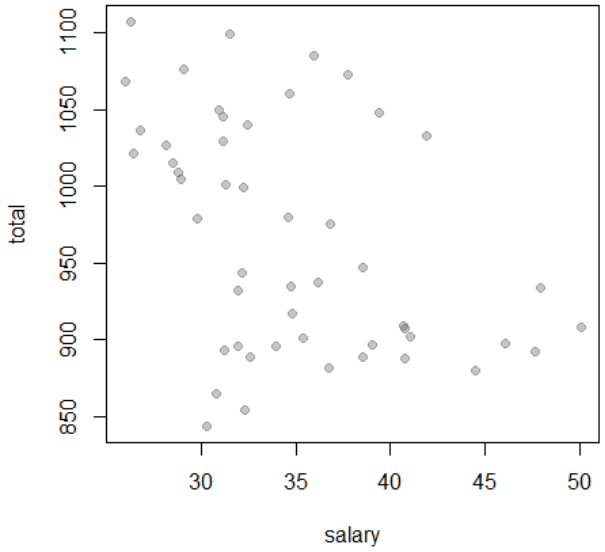
0.802
vs.
0.957

Correlation is not causation

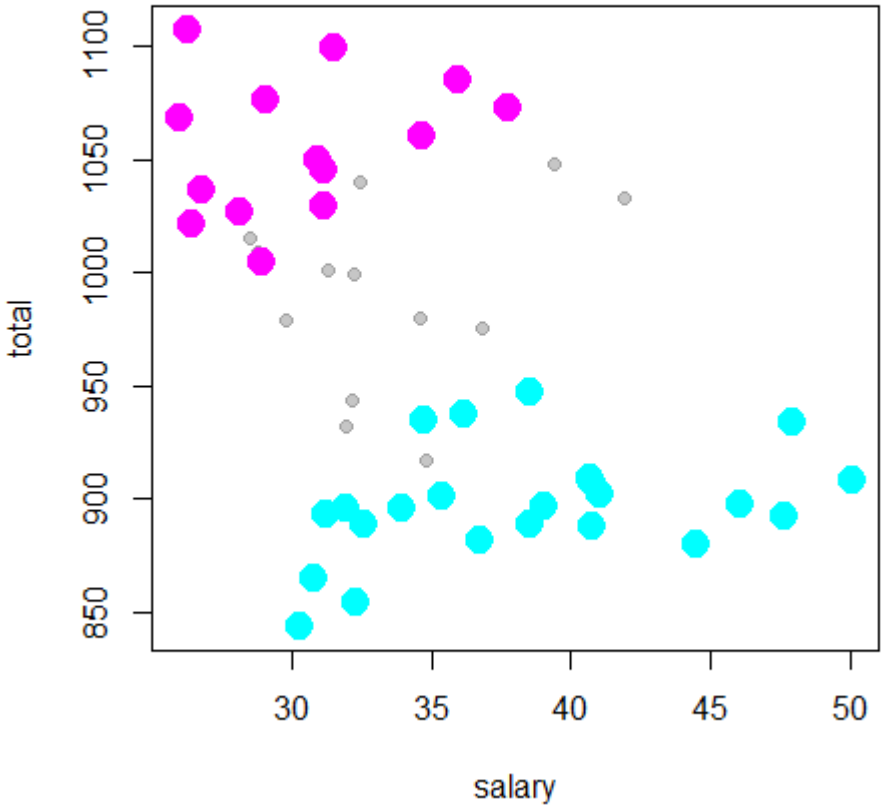
- Lurking variables (compounding effect) – correlates with both response and predictor

ex)

Teacher pay vs. student SAT scores



Correlation is not causation



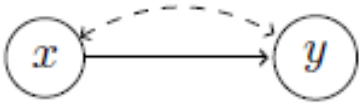
상위 10% 점수 - red
하위 40% 점수 – blue

구간 나누면 양의 상관관계
  상위 10%: 0.2588
  중간: 0.2225
  하위 40%: 0.3673
전체 음의 상관관계
  -0.439

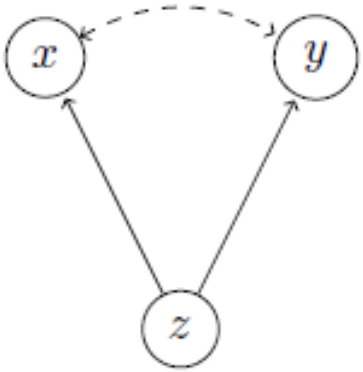
Simpson’s paradox

Association

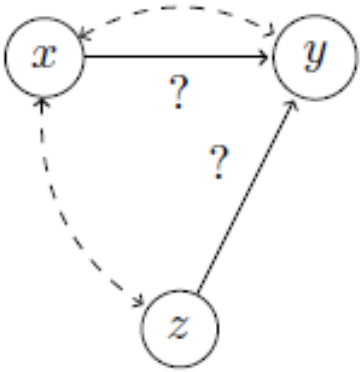
- Association between y (response) and x (explanatory)
- Case of gun ownership and violence
  - Buyback vs. suicide rate (causal?)



(a) Causation



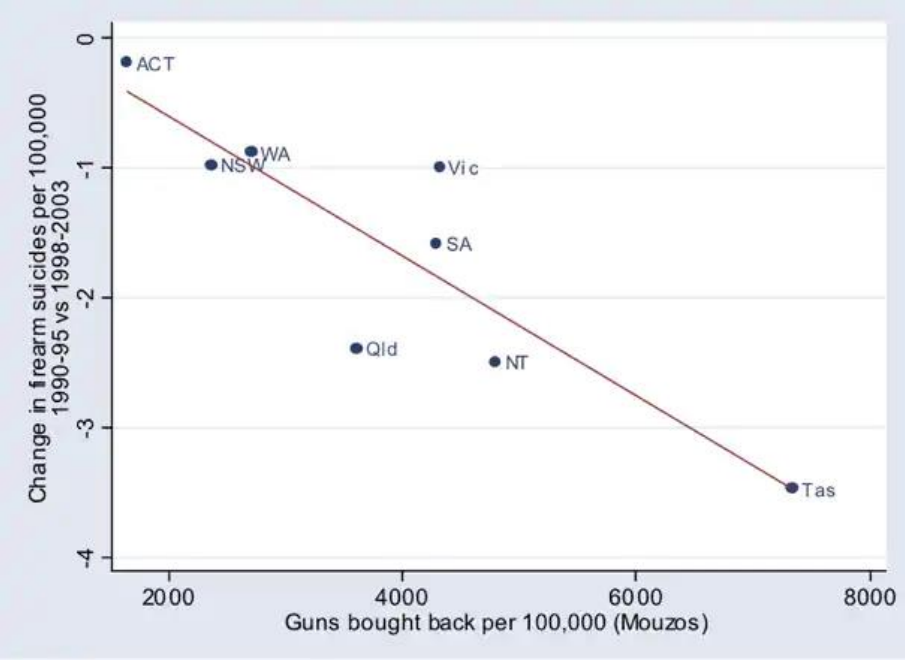
(b) Common response



(c) Confounding

Gun ownership vs. violence

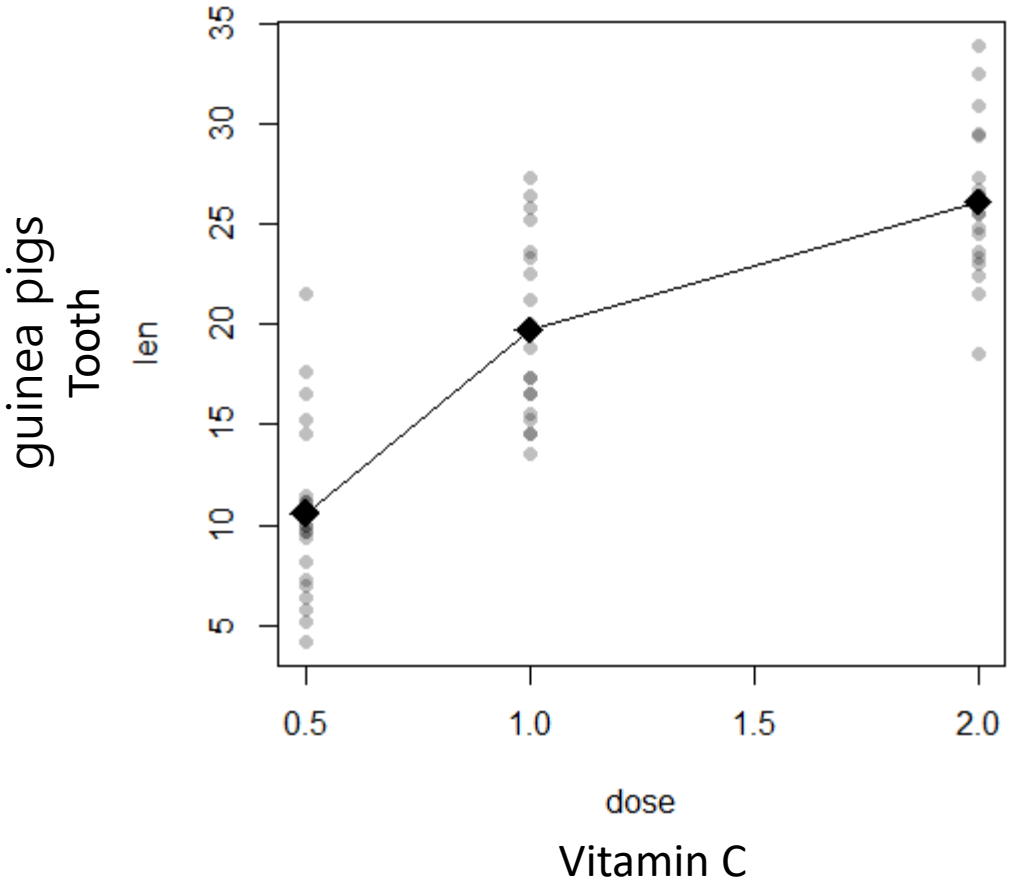
Gun boughtback → Suicide decrease?



A massacre in Tasmania that left 35 dead (1996)

Trends (경향 분석)

- Summary of a relationship between two numeric variables



- Vitamin C가 1단위 증가할 때 Tooth length는 얼마나 증가할 것인가?
- 만약 선형 관계 있다면 하나의 비율 값으로 정의할 수 있음

Trends (경향 분석)

- The **mean response value** depends **linearly** on the predictor value

$$\mu_{y|x} = b_0 + b_1 x$$

$$y_i = b_0 + b_1 x_i + e_i$$

- error term e_i
- A common assumption: the average value of e_i is close to 0

$$\hat{y}_i = b_0 + b_1 x_i$$

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

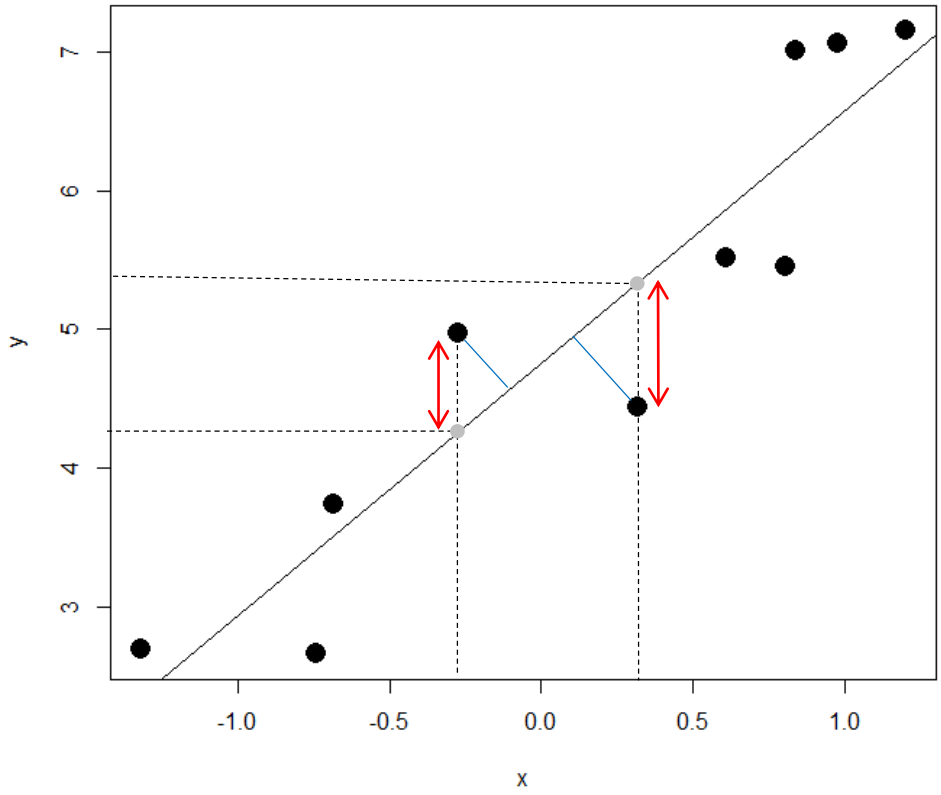
Least square method (최소제곱법)

- Residual

$$\hat{y}_i = b_0 + b_1 x_i$$

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

- Least squares regression line – line which minimizes the squared residuals

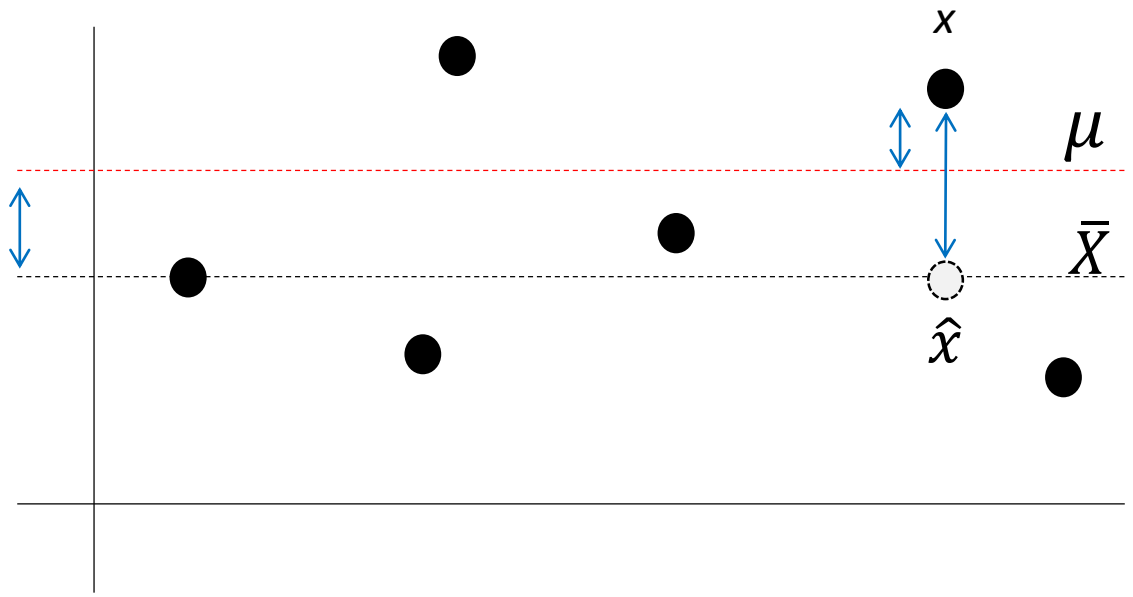


Error vs. Residual

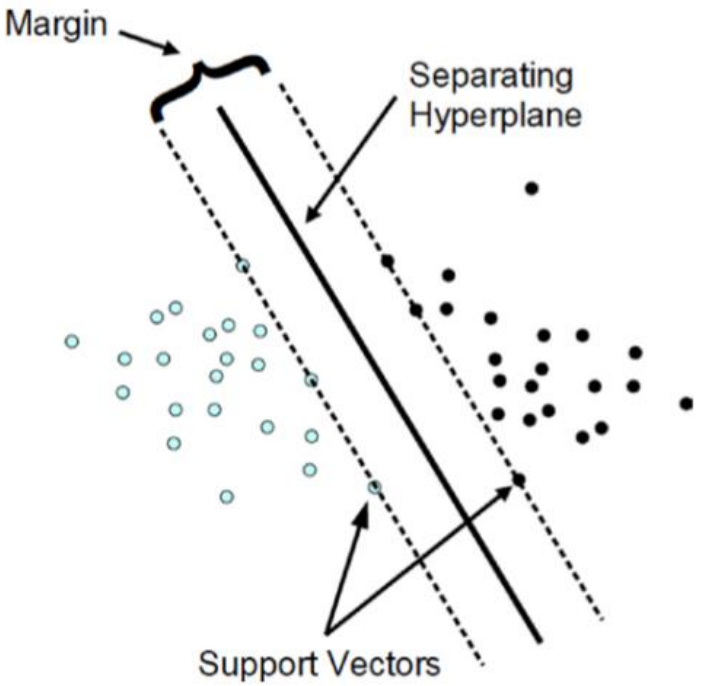
- Error - Diff. between observed data (x) and population statistics (mean) (μ)
- Residual - Diff. between observed data (x) and predicted data ($\hat{x}$)

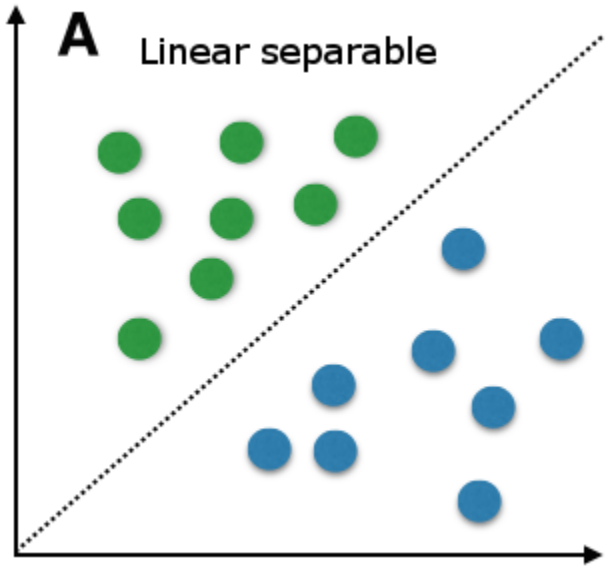
*$\mu \approx$ sample mean ($\bar{X}$)

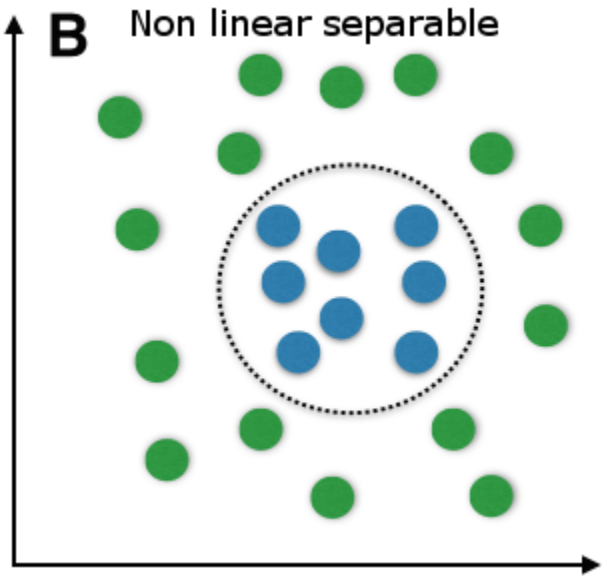
*Predicted data ($\hat{x}$) = model ($\bar{X}$) with observed data(x)



SVM (Machin learning)







Least square method

- Solving the minimization problem
- Fitting a linear model

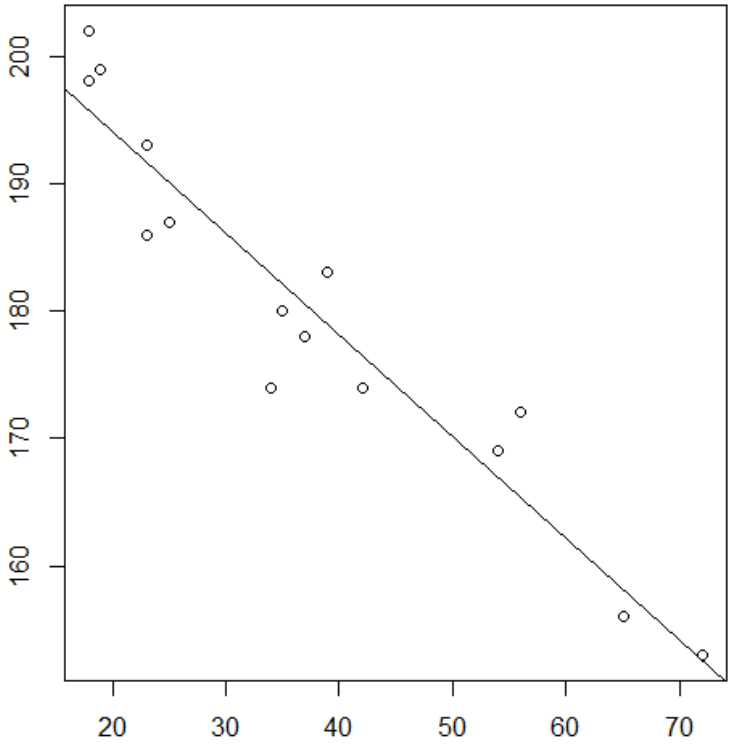
$$\widehat{b_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \text{cor}(x, y)\frac{s_y}{s_x}$$

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

- The slope is related to the correlation
- x, y is not interchangeable
- Sum of the residuals will be 0

Simple regression (단순 회귀)

$$y_i = b_0 + b_1 x_i + e_i$$



- One response variable, one explanatory variable
- Two parameters: b0 and b1
- Error e_i ~ N(0, σ^2), iid (independent and identically distributed)

Exercise)

사용 년수 (year)	1.0	1.5	2.0	2.0	3.0	3.2	4.0	4.5	5.0
중고차 가격 (price)	4.5	4.0	3.2	3.4	2.5	2.3	1.6	1.5	1.0

x = year
y = price

mean(x)=
mean(y)=
sd(x)=
sd(y)=
cor(x, y)=
b1=
b0=
y = b0 + b1*x
산점도 그리기
회귀직선그리기

2. (Paired) Bivariate categorical data

같은 샘플에서 두 변수 (범주형) 측정 비교

- 가족의 규모에 따라 세탁기의 크기가 다른가
- 소득 수준에 따라 주거하는 집의 크기가 다른가

Bivariate categorical data

- Q: Is there a relationship between the variables?
- Two-way contingency tables from summarized data

		Children	
		buckled	unbuckled
Parents	buckled	56	8
	unbuckled	2	16

Seatbelt in California

Marginal distributions

- Marginal distributions of two-way tables

		Children		
		buckled	unbuckled	
Parents	buckled	56	8	64
	unbuckled	2	16	18
		58	24	82

Conditional distribution

Q: whether a parent wearing a seatbelt changes the chance a child does?

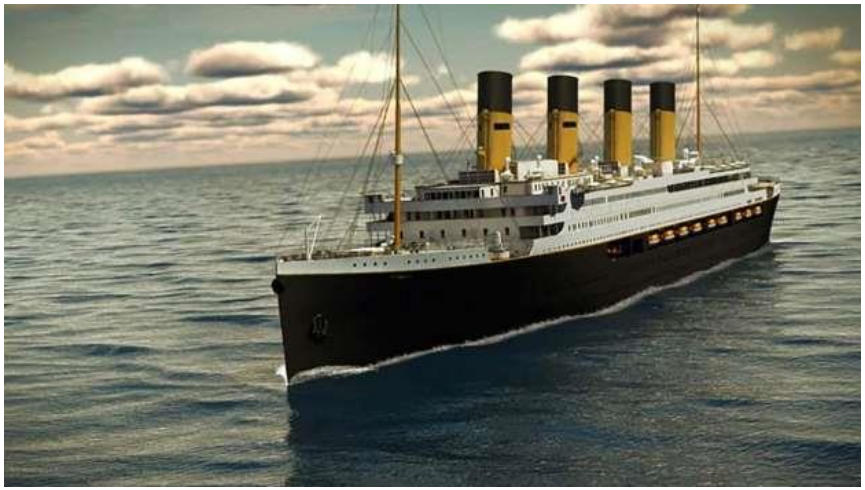
		Children		
		buckled	unbuckled	
Parents	buckled	0.88	0.12	64
	unbuckled	0.11	0.89	18
		58	24	82

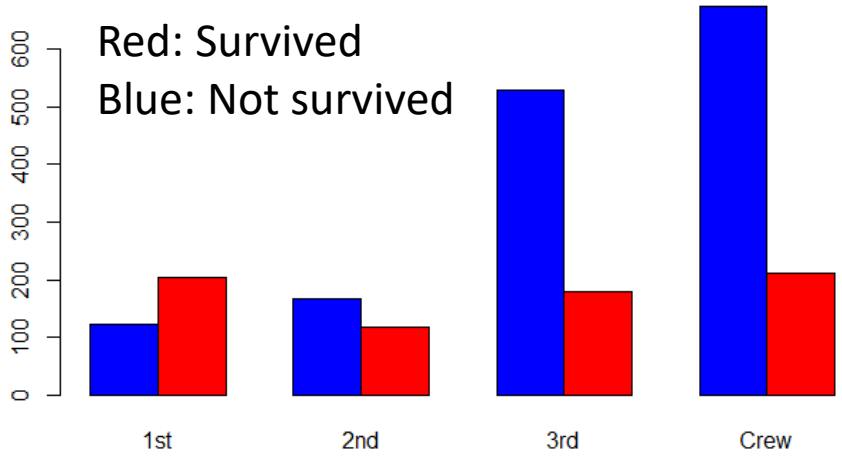
$$p(C = b|P = b) = \frac{p(C = b, P = b)}{p(P = b)}$$

Measures of association

- Kendall tau correlation
- (x1, y1) and (x2, y2) is concordant if x1 and y1 are higher ranked and (x2, y2) is lower ranked

$$\tau = \frac{\#\ of\ concordant\ pair\ - \#\ of\ discordant\ pair}{n(n-1)/2}$$





Measures of association

- The chi-squared statistic – a common summary of a table
- summary function

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o: observed
f_e: expected

Data:
(p1, c1)
(p2, c2)
…
(p82, c82)

f_o: observed (f_e: expected)

		Children		
		buckled	unbuckled	
Parents	buckled	56 ( )	8 ( )	64 ( )
	unbuckled	2 ( )	16 ( )	18 ( )
		58 ( )	24 ( )	82

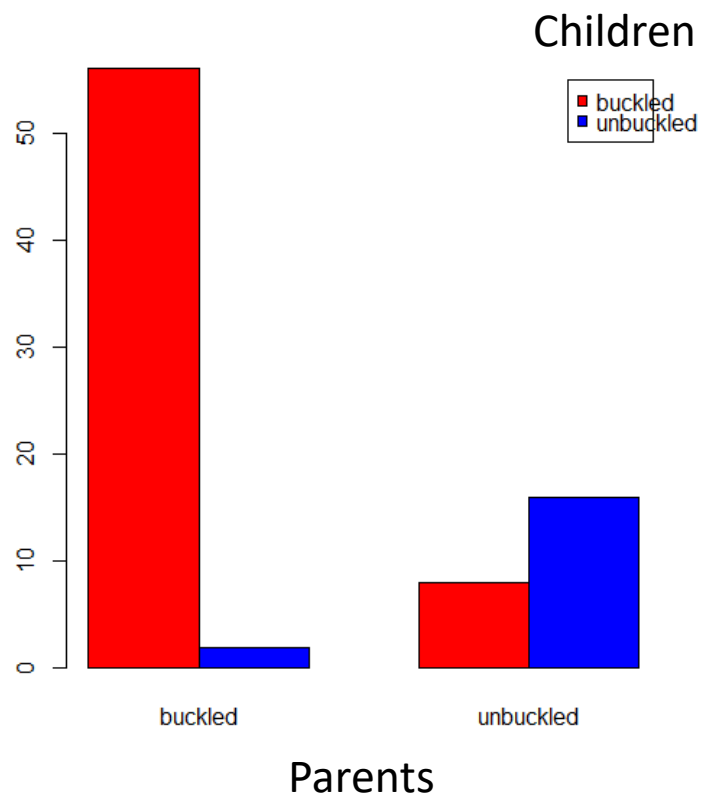
Q: Does the fact that a parent wears a seat belt affect the chance a child does?

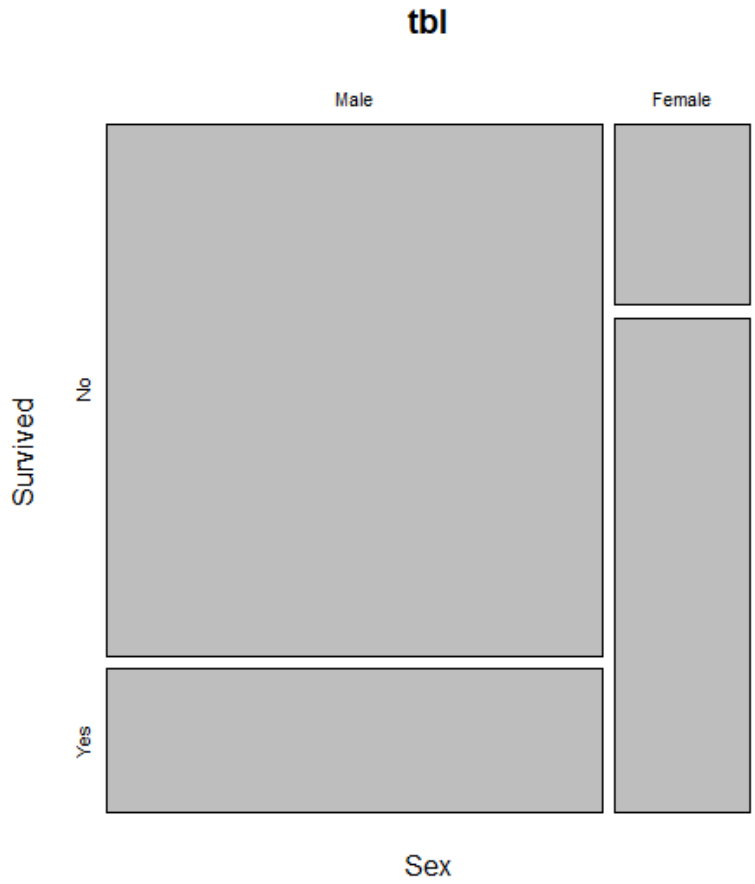
사건 C, P 독립이면 $p(C,P) = p(C) \cdot p(P)$

$N \cdot p(C,P)$ $N \cdot p(C) \cdot p(P)$

Graphical summaries

- barplot
- mosaic plot





숙제 #1 (solution 1)

1. Independent data, 다음 데이터들의 타입을 구분하시오 (명목, 순서, 구간, 비율 중 하나)

직업, 지역, 물가지수, 돈, 성적, 소득, 학년, 혼인 상태, 지지도, 선호도, 몸무게
명 명 구 비 순 비 순 명 순 순 비

2. 다음 데이터셋의 mean, median, variance 를 구하시오

11, 20, 9, 95, 34, 7, 14, 39, 12, 29, 21

mean:26.45, median: 20, variance: 627.67

숙제 #1 (solution 2)

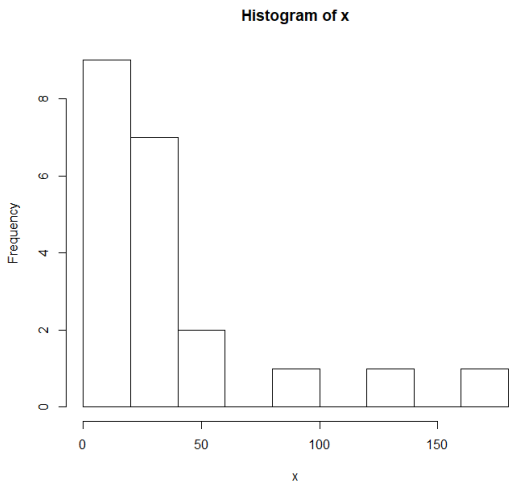
3. 다음 데이터셋의 histogram을 그리시오 (0부터 150까지 10개 구간)

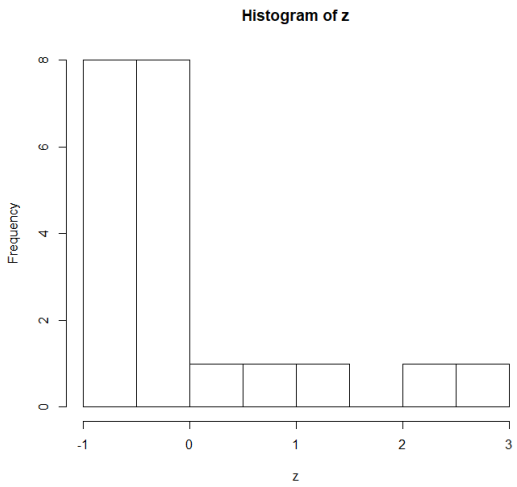
21, 60, 35, 17, 36, 29, 162, 88, 31, 6, 135, 13, 20, 9, 14, 28, 42, 10, 35, 2, 16

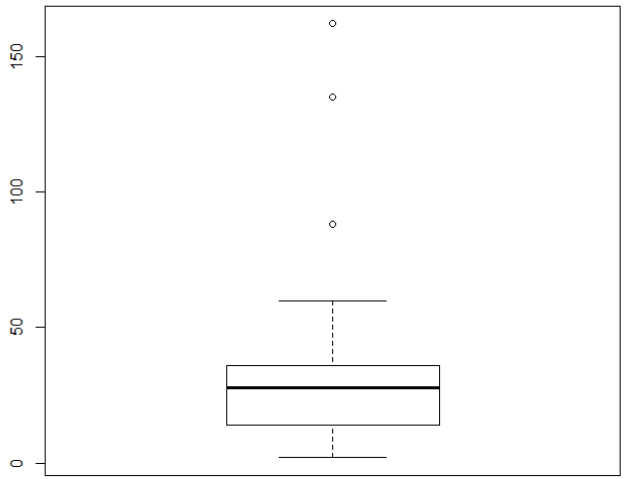
4. 위 예제3의 데이터를 z-score로 변환하고 -1부터 3까지 10개 구간으로 나누어 histogram을 그리시오

5. 위 예제3 데이터의 boxplot을 그리시오 (outlier 무시)

[min] 2
[1Q] 14
[2Q] 28
[3Q] 36
[max] 60







숙제 #2 (다음시간제출, A4용지 사용, 이름, 학번 명시)

1. 다음 X와 Y로부터 두 변수의 correlation을 계산하고 그 의미를 해석하시오

X	1	2	3	4	5
Y	7	5	3	1	-1

2. 어느 화학 제품의 공정 수율(Y)을 그 제품을 만들 때 들어가는 원료의 촉매량(X)에 영향을 받는 것으로 알려져 있다. 그 관련성을 알기 위해 다음 데이터를 얻었다. 설명변수 X와 반응 변수 Y 사이에 회귀직선을 적합하고 산점도를 그린 후 회귀직선을 그리시오.

X	3.5	3.9	3.2	4.2	4.8	3.0	3.2	3.5	2.9	3.8
Y	80.5	85.5	83.5	90.5	92.4	79.8	78.5	84.5	87.2	90.0

3. 한 의학연구가에 의하면 흡연은 눈가에 주름이 지게 하는 요인이 된다고 한다. 이러한 주장이 타당한가를 알아보기 위해 30대 남자 1000명을 랜덤하게 추출하여 조사한 결과 다음과 같은 표를 얻었다. 30대 남자들을 대상으로 볼 때 연구가의 주장이 옳은지 판단하는 연관성을 나타내는 카이제곱 값을 구하라.

	주름있음	주름없음
흡연자	186	114
비흡연자	228	472

Next

Multivariate data Population

- Random variables
- Density functions
- Families of distributions