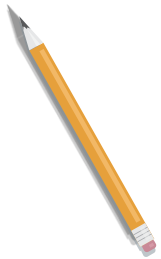
2강

보건정보 데이터의 요약

정보통계학과

이태림 교수



목차

학습하기

1

1. 보건정보 데이터의 분석절차

보건통계학(Public Health Statistics)

인간의 생명현상과 건강문제를 집단적인 현상으로 파악하고 분석하여 국민의 건강증진에 도움을 주는 연구나 사업의 길잡이 역할을 하는 통계적 방법론

보건통계 관련 학회

KOSHIS

IBS

생명과학 자료

생물통계학(Biostatistics)

의약학, 보건학, 역학, 생물학 등의 생명과학 연구분야에서 연구의 계획과 설계,

데이터의 수집과정 및 분석과 해석에 이르는 전 과정에 필요한 통계학적 방법론

생명과학 자료분석 분야

- 인구학(demography): 출생, 사망, 인구추이
- 임상연구(clinical trials): 진단, 치료비교, 생존모형
- 환경통계(Environmental Statistics): 공기오염모형
- 유전통계(genomics): 암 유전요인, 혈우병 유전요인
- 역학통계(Epidemiological statistics): 질병원인추정모형
- 생물통계(Biometrics): 미생물 분포모형

IBC(International Biometric Conference) 2020 5-10 July

연구과정

보건정보 데이터 분석과정

계획
표본수 결정 → 설계
표본추출 방법 → 데이터 수집
표본추출 방법 → 모수 추정 → 가설검정 → 발표

기본개념

- 모집단(Population): 관심의 대상이 되는 전체
- 표본(Sample): 모집단을 대표하는 일부분
- 모수(Parameter): 모집단의 특징을 수치로 표현한 것
- 통계량(Statistics): 표본정보를 양적으로 축약한 표본함수값

학습하기

1 2

2. 보건정보 데이터의 형태

보건자료의 예

치아우식증 자료

```
 1. hilo    response var.        increment=0  noincrement=1
 2. sex                          male=1    Female=2
 3. residence floridation        resident=1 nonresidence=2
 4. fluorisis                    + = 1      - = 2
 5. born.....born in fluoridation area     + = 1    - = 2
 6. water-f      fluoridation water      use = 1     nouse = 2
 7. sib                          + = 1    - = 2
 8. occupation of mother         + = 1    - = 2
 9. visit      Dental visits                1 : nonuse
                                            2 : some year
                                            3 : occasionally
                                            4 : 1 time per 1 year
10. wk Frequency of brushing/week          1 : no brushing
                                            2 : 1 time/1 week
                                            3 : 1 time/2-3days
                                            4 : 1 time/1 day
11. snack  Frequency of snacks             1 : no snack
                                            2 : 1 snack/2-3days
                                            3 : 1 snack/1 day
12. sugar              0 : low sugar     1 : high sugar
13. q-snack    frequency of quick snack
14. oral hygiene       0, 1, 2, 3   level of oral hygiene
15. lacto
16. smutan              0 = none ,    1 = 10^4 , 2 = 5×10^4 ,
                        3 = 2.5×10^5 , 4 = 5×10^5/mlsaliva
17. dst1     # of decayed surface at 6 years
18. dss1     #  of smooth surface at 6 years
19. dsp1     # of pit and fissure at 6 years
20. whp      # of white spot at 6 years
21. mor      # of   morphology scores
```

보건자료의 형태

양적자료

치아우식증 자료

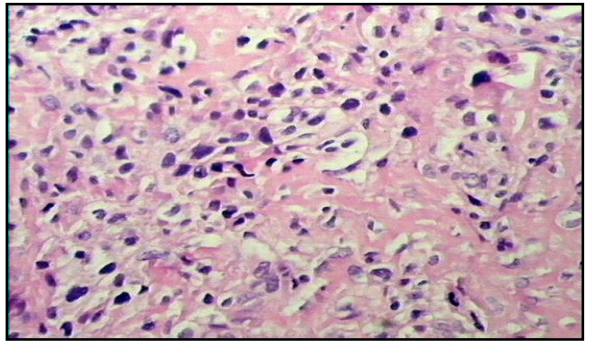
질적자료

1. hilo	response var.	increment=0 noincrement=1
2. sex		male=1 Female=2
3. residence floridation		residence=1 nonresidence=2
4. fluorisis		+ = 1 − = 2
5. born	born in fluoridation area	+ = 1 − = 2
6. water-f	fluoridation water	use = 1 nouse = 2
7. sib		+ = 1 − = 2
8. occupation of mother		+ = 1 − = 2
9. visit	Dental visits	1 : nonuse
		2 : some year
		3 : occasionally
		4 : 1 time per 1 year
10. wk	Frequency of brushing/week	1 : no brushing
		2 : 1 time/1 week
		3 : 1 time/2-3days
		4 : 1 time/1 day
11. snack	Frequency of snacks	1 : no snack
		2 : 1 snack/2-3days
		3 : 1 snack/1 day
12. sugar	0 : low sugar	1 : high sugar
13. q-snack	frequency of quick snack	
14. oral hygiene	0, 1, 2, 3	level of oral hygiene
15. lacto		
16. smutan		0 = none , 1 = 10^4 , 2 = 5×10^4 , 3 = 2.5×10^5 , 4 = 5×10^5/misaliva
17. dst1	# of decayed surface at 6 years	
18. dss1	# of smooth surface at 6 years	
19. dsp1	# of pit and fissure at 6 years	
20. wsp	# of white spot at 6 years	
21. mor	# of morphology scores	

보건자료의 형태

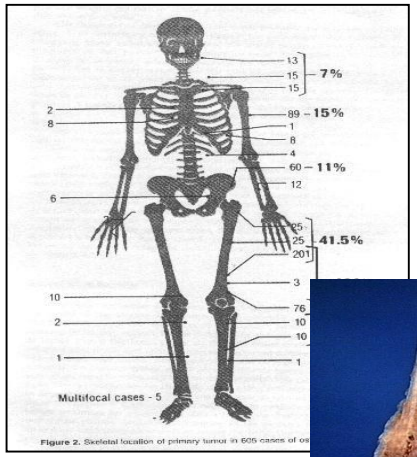
양적자료

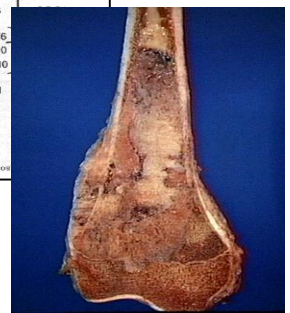
골수암 자료



질적자료

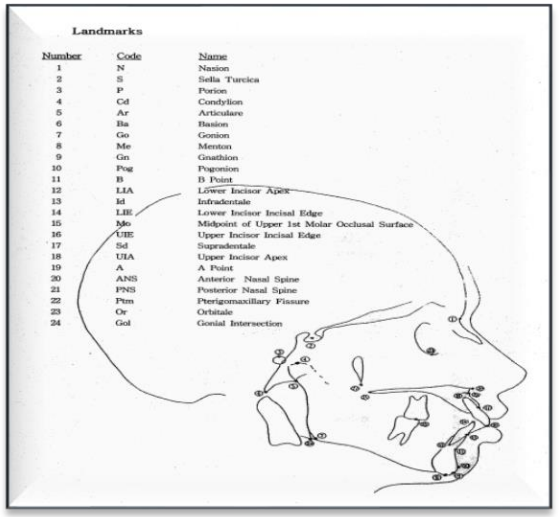
Characteristic	Number of Paients	%
Age		
<12	59	16.2
≥12 and ≤15	110	30.1
>15 and ≤40	176	48.2
>40	20	5.5
Gender		
Male	241	66.0
Female	124	34.0
Maximal tumor diameter		
≤6cm	65	17.8
>6cm and ≤8	76	20.8
>8cm	224	61.4
Subtype		
Osteoblastic	305	83.6
Chondroblastic	28	7.7
Fibroblastic	14	3.8
Other	18	4.9
Location		
Distal femur	182	49.9
Proximal tibia	88	24.1
Proximal humerus	34	9.3
Fibula	29	7.9
Other	31	8.8
Pathologic fracture		
Absent	347	95.1
Present	18	4.9
Tumor necrotic rate		
100%	59	16.2
≥90% and <100%	99	27.1
≥50% and <90%	85	23.3
<50%	122	33.4

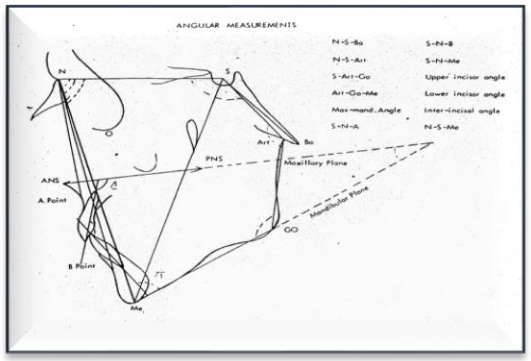


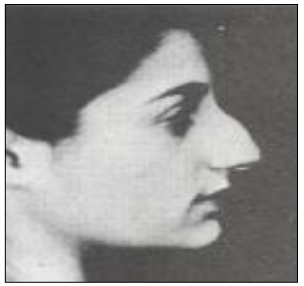


보건자료의 형태

안악면 성장 자료







양적자료

질적자료

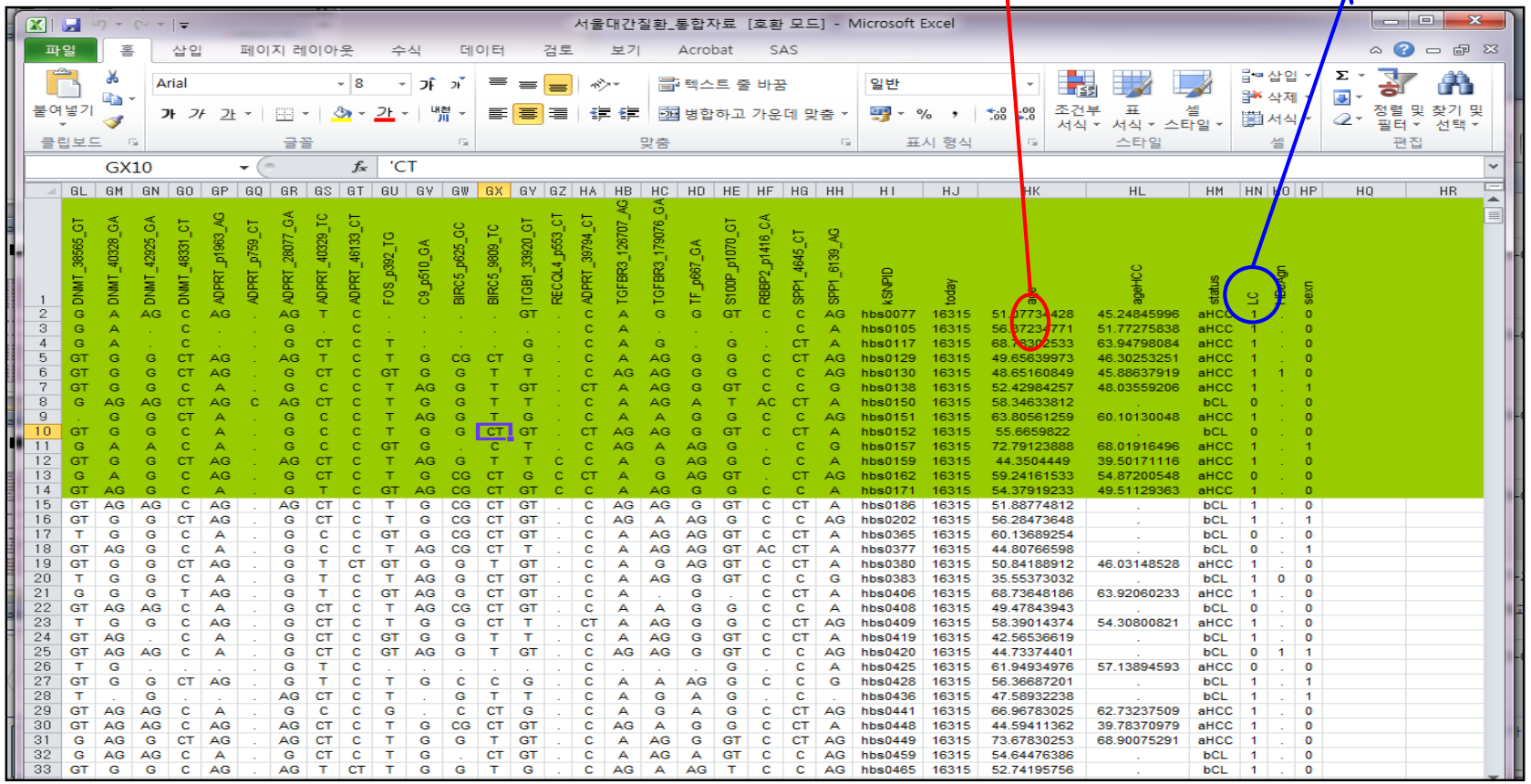
- Data : 675 lateral cephalograms of 6 years old children consisted 337 males and 338 females
  Variable : 136 variables which consisted of 54 angular measurements and 82 linear measurements from the X , Y coordinates of 24 landmarks

보건정보학 Health Informatics

간암 자료

양적자료

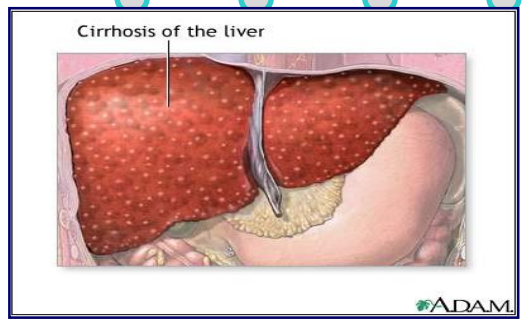
질적자료

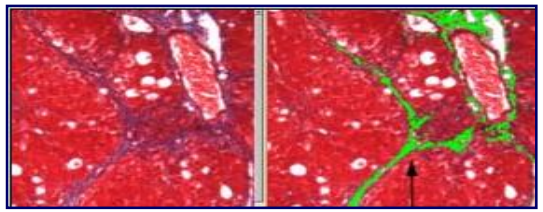


서울대간질환_통합자료 [호환 모드] - Microsoft Excel

GX10 'CT

	GL	GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV	GW	GX	GY	GZ	HA	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK	HL	HM	HN	HO	HP	HQ	HR
	DNMT_38565_GT	DNMT_40328_GA	DNMT_42925_GA	DNMT_48331_CT	ADPRT_p1963_AG	ADPRT_p759_CT	ADPRT_28077_GA	ADPRT_40329_TC	ADPRT_46133_CT	FOS_p392_TG	C9_p510_GA	BIRC5_p625_GC	BIRC5_9809_TC	ITGB1_33920_GT	RECQL4_p553_CT	ADPRT_39794_CT	TGFBR3_126707_AG	TGFBR3_179076_GA	TF_p667_GA	S100P_p1070_GT	RBBP2_p1416_CA	SPP1_4645_CT	SPP1_6139_AG	kSNPID	today	age	ageHCC	status	LC	HBsAg	sex		
1																																	
2	G	A	AG	C	AG	.	AG	T	C	.	.	.	.	GT	.	C	A	G	G	GT	C	C	AG	hbs0077	16315	51.07734428	45.24845996	aHCC	1	.	0		
3	G	A	.	C	.	.	G	.	C	.	.	.	.	.	.	C	A	.	.	.	.	C	A	hbs0105	16315	56.37234771	51.77275838	aHCC	1	.	0		
4	G	A	.	C	.	.	G	CT	C	T	.	.	.	G	.	C	A	G	.	G	.	CT	A	hbs0117	16315	68.78302533	63.94798084	aHCC	1	.	0		
5	GT	G	G	CT	AG	.	AG	T	C	T	G	CG	CT	G	.	C	A	AG	G	G	C	CT	AG	hbs0129	16315	49.65639973	46.30253251	aHCC	1	.	0		
6	GT	G	G	CT	AG	.	G	CT	C	GT	G	G	T	T	.	C	AG	AG	G	G	C	C	AG	hbs0130	16315	48.65160849	45.88637919	aHCC	1	1	0		
7	GT	G	G	C	A	.	G	C	C	T	AG	G	T	GT	.	CT	A	AG	G	GT	C	C	G	hbs0138	16315	52.42984257	48.03559206	aHCC	1	.	1		
8	G	AG	AG	CT	AG	C	AG	CT	C	T	G	G	T	T	.	C	A	AG	A	T	AC	CT	A	hbs0150	16315	58.34633812	.	bCL	0	.	0		
9	.	G	G	CT	A	.	G	C	C	T	AG	G	T	G	.	C	A	A	G	G	C	C	AG	hbs0151	16315	63.80561259	60.10130048	aHCC	1	.	0		
10	GT	G	G	C	A	.	G	C	C	T	G	G	CT	GT	.	CT	AG	AG	G	GT	C	CT	A	hbs0152	16315	55.6659822	.	bCL	0	.	0		
11	G	A	A	C	A	.	G	C	C	GT	G	.	C	T	.	C	AG	A	AG	G	.	C	G	hbs0157	16315	72.79123888	68.01916496	aHCC	1	.	1		
12	GT	G	G	CT	AG	.	AG	CT	C	T	AG	G	T	T	C	C	A	G	AG	G	C	C	A	hbs0159	16315	44.3504449	39.50171116	aHCC	1	.	0		
13	G	A	G	C	AG	.	G	CT	C	T	G	CG	CT	G	C	CT	A	G	AG	GT	.	CT	AG	hbs0162	16315	59.24161533	54.87200548	aHCC	0	.	0		
14	GT	AG	G	C	A	.	G	T	C	GT	AG	CG	CT	GT	C	C	A	AG	G	G	C	C	A	hbs0171	16315	54.37919233	49.51129363	aHCC	1	.	0		
15	GT	AG	AG	C	AG	.	AG	CT	C	T	G	CG	CT	GT	.	C	AG	AG	G	GT	C	CT	A	hbs0186	16315	51.88774812	.	bCL	1	.	0		
16	GT	G	G	CT	AG	.	G	CT	C	T	G	CG	CT	GT	.	C	AG	A	AG	G	C	C	AG	hbs0202	16315	56.28473648	.	bCL	1	.	1		
17	T	G	G	C	A	.	G	C	C	GT	G	CG	CT	GT	.	C	A	AG	AG	GT	C	CT	A	hbs0365	16315	60.13689254	.	bCL	0	.	0		
18	GT	AG	G	C	A	.	G	C	C	T	AG	CG	CT	T	.	C	A	AG	AG	GT	AC	CT	A	hbs0377	16315	44.80766598	.	bCL	0	.	1		
19	GT	G	G	CT	AG	.	G	T	CT	GT	G	G	T	GT	.	C	A	G	AG	GT	C	CT	A	hbs0380	16315	50.84188912	46.03148528	aHCC	1	.	0		
20	T	G	G	C	A	.	G	T	C	T	AG	G	CT	GT	.	C	A	AG	G	GT	C	C	G	hbs0383	16315	35.55373032	.	bCL	1	0	0		
21	G	G	G	T	AG	.	G	T	C	GT	AG	G	CT	GT	.	C	A	.	G	.	C	CT	A	hbs0406	16315	68.73648186	63.92060233	aHCC	1	.	0		
22	GT	AG	AG	C	A	.	G	CT	C	T	AG	CG	CT	GT	.	C	A	A	G	G	C	C	A	hbs0408	16315	49.47843943	.	bCL	0	.	0		
23	T	G	G	C	AG	.	G	CT	C	T	G	G	CT	T	.	CT	A	AG	G	G	C	CT	AG	hbs0409	16315	58.39014374	54.30800821	aHCC	1	.	0		
24	GT	AG	.	C	A	.	G	CT	C	GT	G	G	T	T	.	C	A	AG	G	GT	C	CT	A	hbs0419	16315	42.56536619	.	bCL	1	.	0		
25	GT	AG	AG	C	A	.	G	CT	C	GT	AG	G	T	GT	.	C	AG	AG	G	GT	C	C	AG	hbs0420	16315	44.73374401	.	bCL	0	1	1		
26	T	G	.	.	.	.	G	T	C	.	.	.	.	.	.	C	.	.	.	G	.	C	A	hbs0425	16315	61.94934976	57.13894593	aHCC	0	.	0		
27	GT	G	G	CT	AG	.	G	T	C	T	G	C	C	G	.	C	A	A	AG	G	C	C	G	hbs0428	16315	56.36687201	.	bCL	1	.	1		
28	T	.	G	.	.	.	AG	CT	C	T	.	G	T	T	.	C	A	G	A	G	.	C	.	hbs0436	16315	47.58932238	.	bCL	1	.	1		
29	GT	AG	AG	C	A	.	G	C	C	G	.	C	CT	G	.	C	A	G	A	G	C	CT	AG	hbs0441	16315	66.96783025	62.73237509	aHCC	1	.	0		
30	GT	AG	AG	C	AG	.	AG	CT	C	T	G	CG	CT	GT	.	C	AG	A	G	G	C	CT	A	hbs0448	16315	44.59411362	39.78370979	aHCC	1	.	0		
31	G	AG	G	CT	AG	.	AG	CT	C	T	G	G	T	GT	.	C	A	AG	G	GT	C	CT	AG	hbs0449	16315	73.67830253	68.90075291	aHCC	1	.	0		
32	G	AG	AG	C	A	.	G	CT	C	T	G	.	CT	GT	.	C	A	AG	A	GT	C	C	AG	hbs0459	16315	54.64476386	.	bCL	1	.	0		
33	GT	G	G	C	AG	.	AG	T	CT	T	G	G	T	G	.	C	AG	A	AG	T	C	C	AG	hbs0465	16315	52.74195756	.	bCL	1	.	0		





학습하기

2 3

3. 자료의 기술 및 요약

질적자료의 기술

- 분할표(contingency Table): 특성에 의해 분류한 표

  상대위험률(RR), 오즈비(Odds Ratio)

아스피린 복용과 심근경색 발생간의 관계

	심근경색	
	유	무
아스피린	139	10898
플라시보	239	10795

상대위험률(Relative Risk)

아스피린 복용과 심근경색 발생간의 관계

	심근경색		
	유	무	계
아스피린	139	10898	11037
플라시보	239	10795	11034
계	378	21693	

(출처: Dawson-Saunders and Trapp, 1994)

$$RR = \frac{139/11037}{239/11034} = \frac{0.0126}{0.0217} = 0.581$$

상대위험률(RR)이 1 보다 작으므로 아스피린을 복용했을 경우 안 복용했을 때보다 심근경색을 일으킬 위험이 상대적으로 적다는 의미

상대위험률(Relative Risk)

위험요인에 노출된 집단과 그렇지 않은 집단을 일정기간 관찰하여 질병의 발생 또는 사망률을 산출하여 얻은 두 비율의 비

$$RR = \frac{\text{위험군에서의 위험률}}{\text{대조군에서의 위험률}} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$$

코호트연구(Cohort study) 또는 위험인자군과 대조군이 사전에 정해지고 난 후 그 결과를 관찰하게 되는 임상시험연구에서 구할 수 있다.

오즈비(Odds Ratio)

약물남용과 심장발작 발생간의 관계

		심장발작	
		유	무
약물남용	유	73	18
	무	141	196
	계	214	214
	Odds	0.518	0.092

- 심장발작군의 오즈

$$\text{odds 1} = \frac{73/214}{141/214} = 0.518$$

- 정상군의 오즈

$$\text{odds 2} = \frac{18/214}{196/214} = 0.092$$

위험요인과 질병발생 간의 연관성을

1을 기준으로 나타낸 척도

오즈비(Odds Ratio)

약물남용과 심장발작 발생간의 관계

		심장발작	
		유	무
약물남용	유	73	18
	무	141	196
	계	214	214
	Odds	0.518	0.092

오즈비

$$OR = \frac{[n_{11}/n_{+1}]/[n_{21}/n_{+1}]}{[n_{12}/n_{+2}]/[n_{22}/n_{+2}]} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

- 오즈비 계산

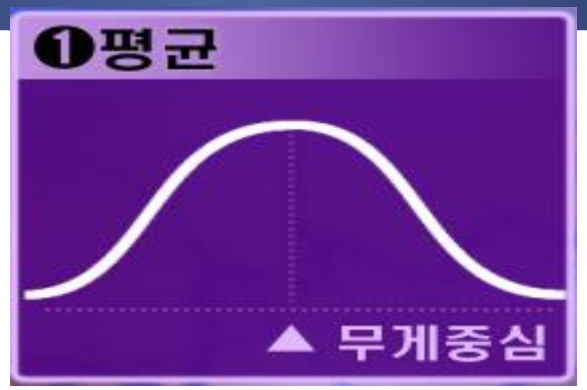
$$OR = \frac{0.518}{0.092} = 5.63$$

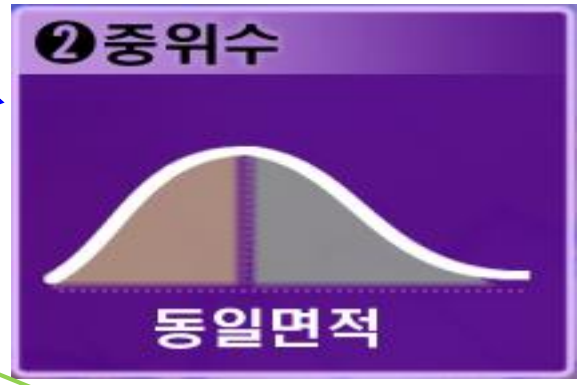
→ 심장발작군에서의 약물남용이 정상군에 비해 5.6배 많다

양적자료의 기술

중심위치

- 평균
- 중위수
- 최빈값
- 사분위수
- 백분위수

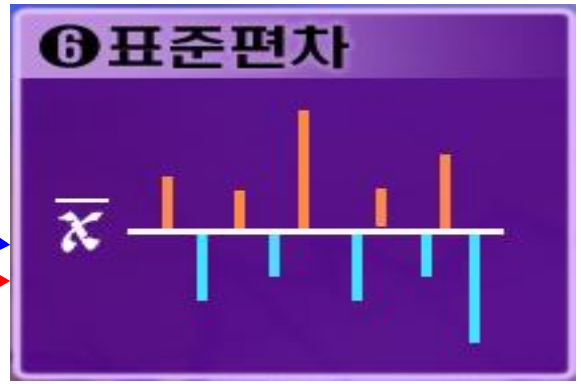


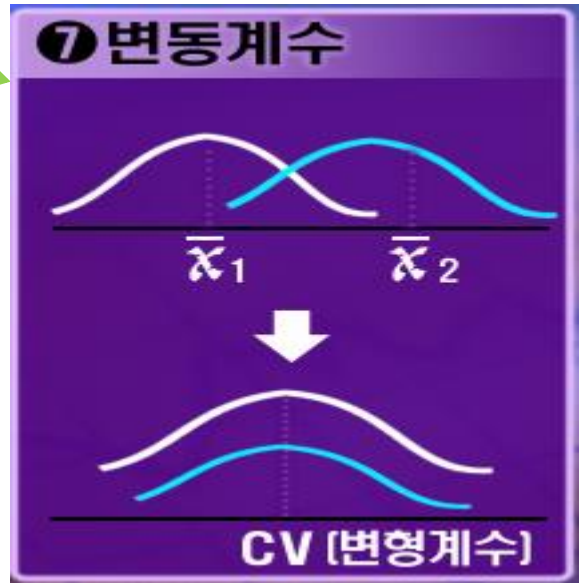


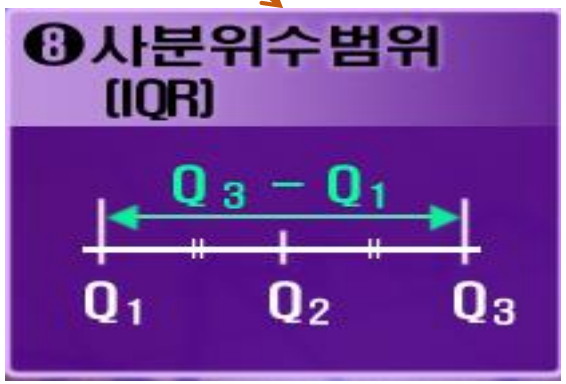
양적자료의 기술

퍼짐정도

- 분산(Variance)
- 표준편차 (Standard Deviation)
- 변동계수 (Coefficient of Variation)
- 사분위수(IQR)



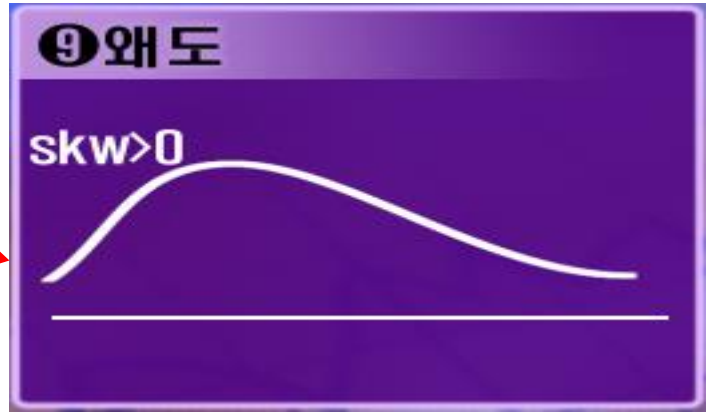


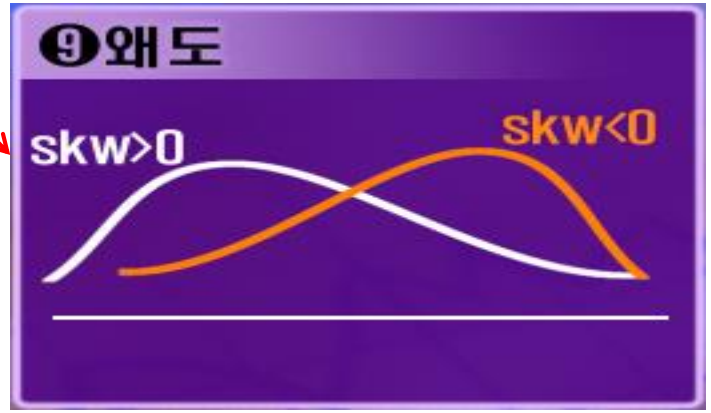


양적자료의 기술

분포특징

- 왜도(skewness)
- 첨도(kurtosis)

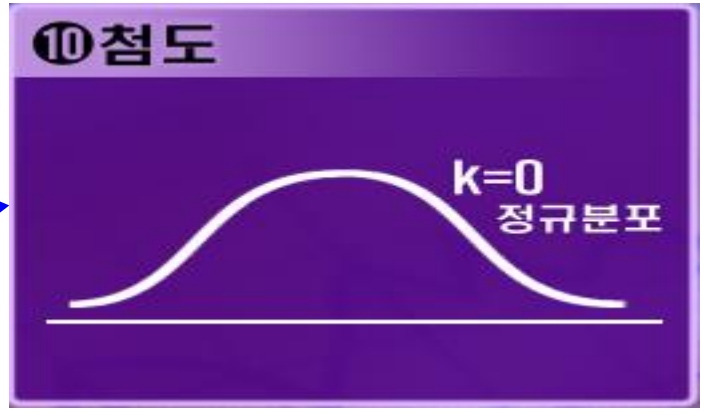


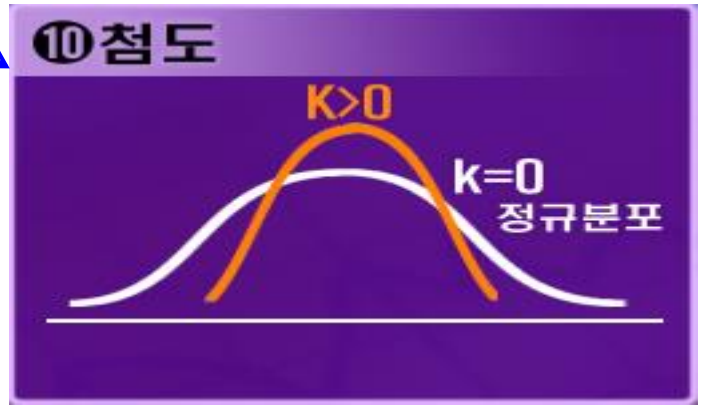


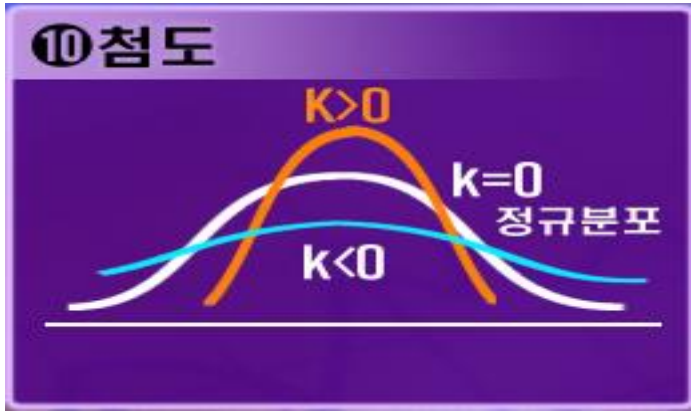
양적자료의 기술

분포특징

- 왜도(skewness)
- 첨도(kurtosis)



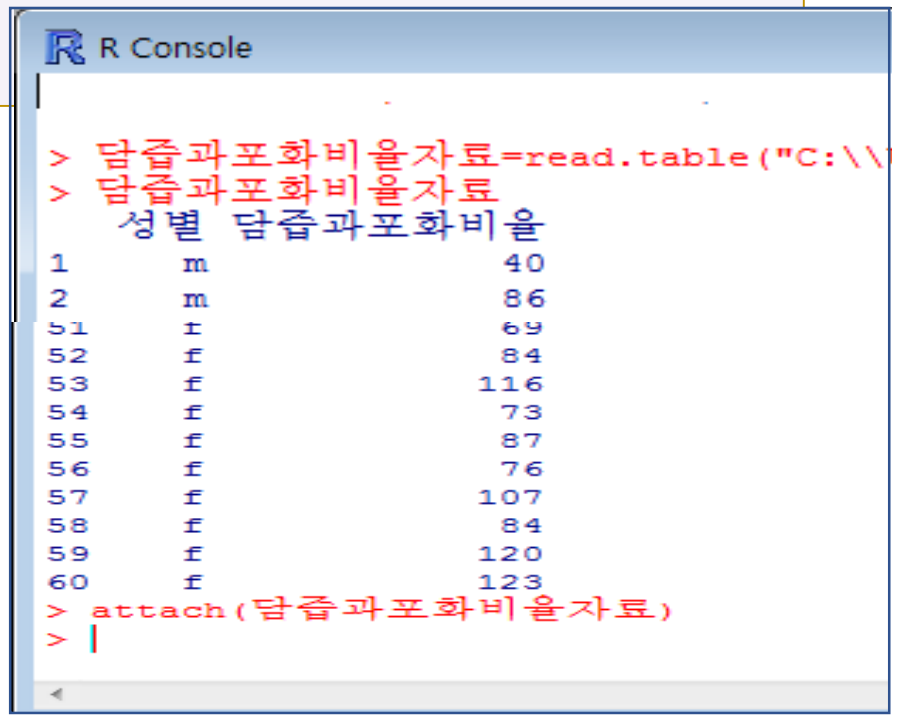




데이터파일 읽기

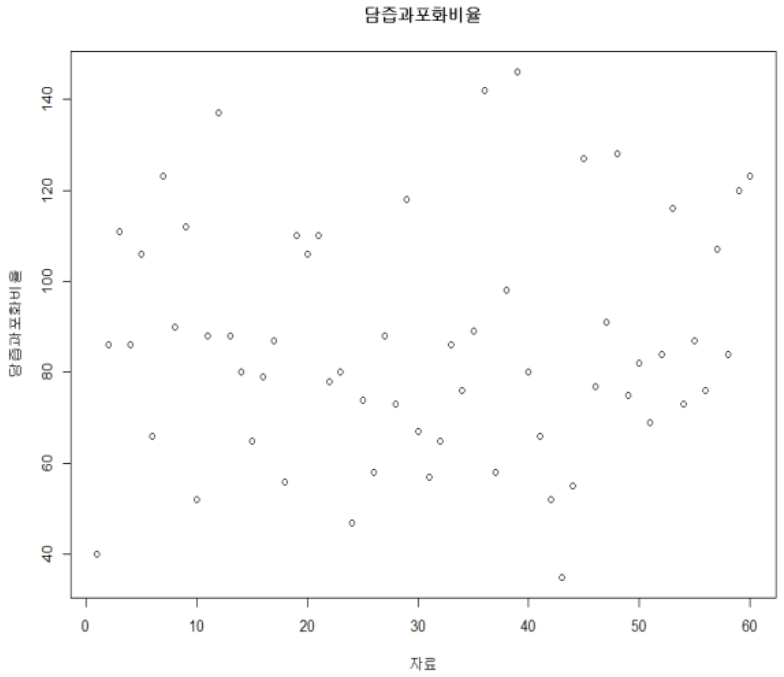
교재 p46 예제 2.3

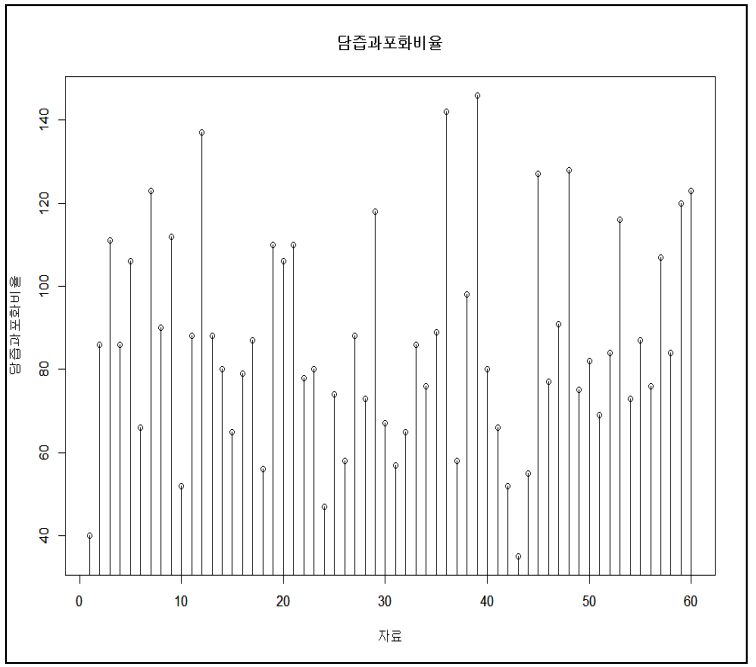
```
담즙과포화비율자료
=read.table("C:₩₩Users₩₩knou₩₩Desktop₩₩R
실습
                자료₩₩담즙과포화비율자료.txt",header=T)
담즙과포화비율자료
attach(담즙과포화비율자료)
```



자료를 순서대로 그림으로 나타내기

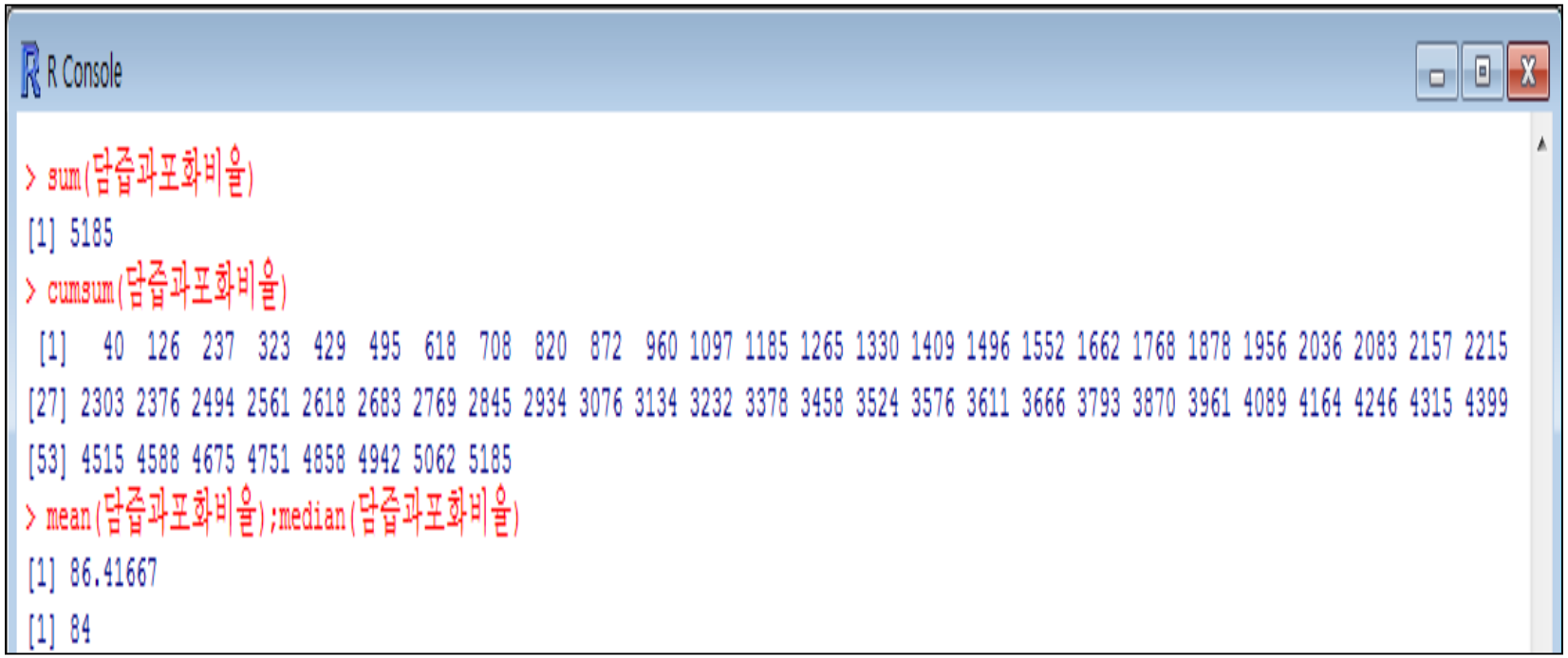

```
>plot(담즙과포화비율, type="p", xlab="자료", ylab="
  담즙과포화비율", main="담즙과포화비율")
>par(new=T)
>plot(담즙과포화비율, type="h", xlab="자료",  ylab="
  담즙과포화비율", main="담즙과포화비율")
```





수치적 요약(중심경향, 흩어진 정도)


```
#합
>sum(담즙과포화비율)
#누적합(cumulative sum)
>cumsum(담즙과포화비율)
#산술평균과 중앙값
>mean(담즙과포화비율);median(담즙과포화비율)
```



수치적 요약(중심경향, 흩어진 정도) 및 그림

#10%절삭평균

>mean(담즙과포화비율, trim=1/10)

#분산과 표준편차

R Console

```
> #10%절삭평균
> mean(담즙과포화비율, trim=1/10)
[1] 85.4375
> #분산과 표준편차
> var(담즙과포화비율);sd(담즙과포화비율)
[1] 657.1624
[1] 25.63518
> #다섯숫자 요약(최소값, 제1사분위수, 중앙값, 제 3사분위수, 최대값)
> quantile(담즙과포화비율)
    0%    25%    50%    75%   100%
 35.00  68.50  84.00 106.25 146.00
> #수치적측도 요약(다섯숫자 요약+산술평균)
> summary(담즙과포화비율)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  35.00   68.50   84.00   86.42  106.20  146.00
```

수치적 요약(중심경향, 흩어진 정도) 및 그림

#사분위수간 범위=제 3사분위수-제 1사분위수

>IQR(담즙과포화비율)

#MAD(Median Absolute Deviation): 각 데이터에서 중앙값을 뺀 후 절대값을 취한 값들의 중앙값

>mad(담즙과포화비율)

#범위

>ran

최대

>Ra

>Ra

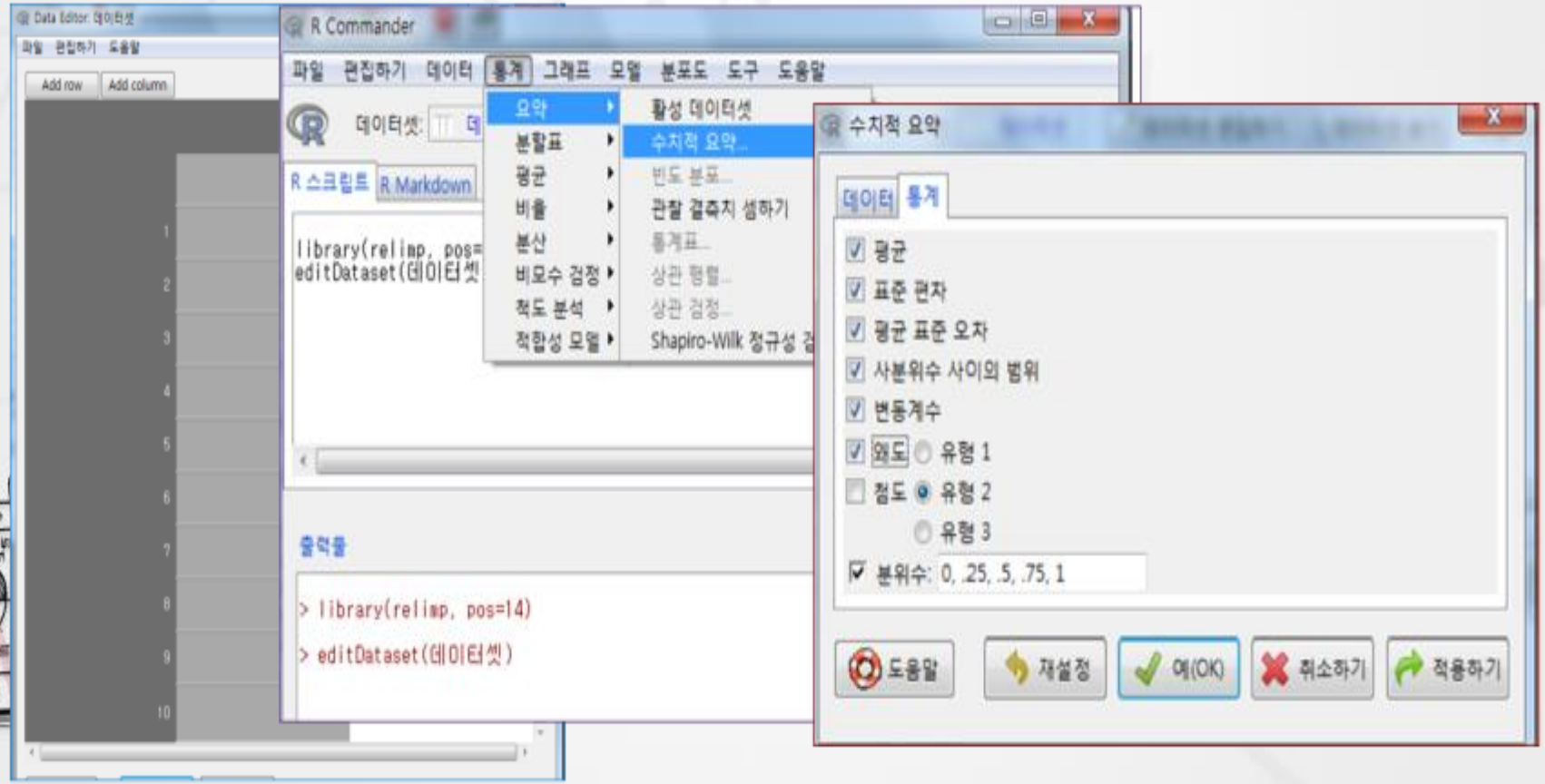
R Console

```
> #사분위수간범위=제 3사분위수-제 1사분위수
> IQR(담즙과포화비율)
[1] 37.75
> #MAD(Median Absolute Deviation): 각 데이터에서 중앙값을 뺀 후 절대값을 취한 값들의 중앙값
> mad(담즙과포화비율)
[1] 26.6868
> #범위
> range(담즙과포화비율) #R에서는 범위를 호출하면 최소값과 최대값을 출력한다
[1]  35 146
> Range=max(담즙과포화비율)-min(담즙과포화비율)
> Range
[1] 111
```

R-commander에 의한 수행

4.3 3.4 3.8 5.2 4.4 2.9 3.7 5.6 4.1 3.6

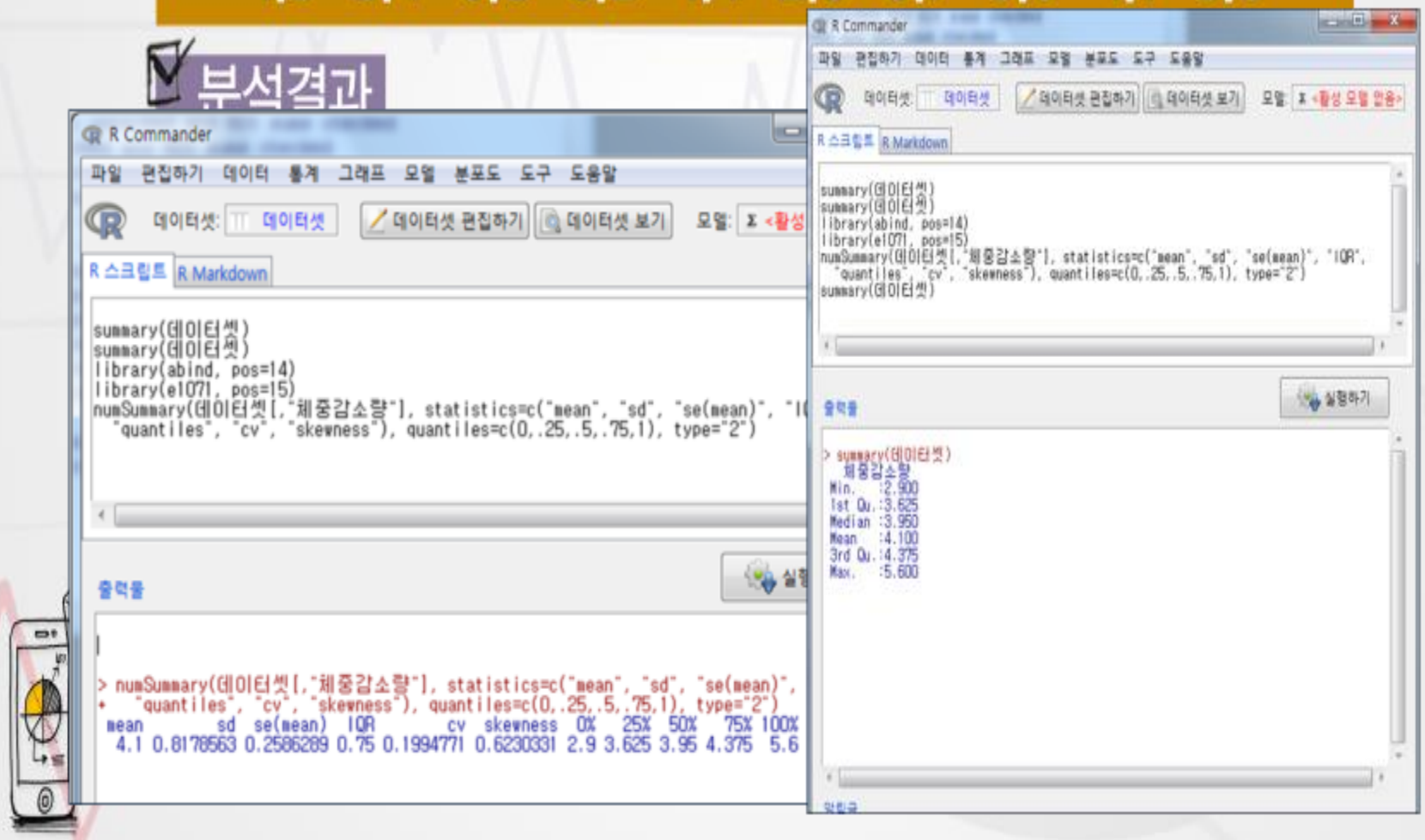
수행과정



식이요법을 실시한 10명의 체중감소량

4.3 3.4 3.8 5.2 4.4 2.9 3.7 5.6 4.1 3.6

분석결과

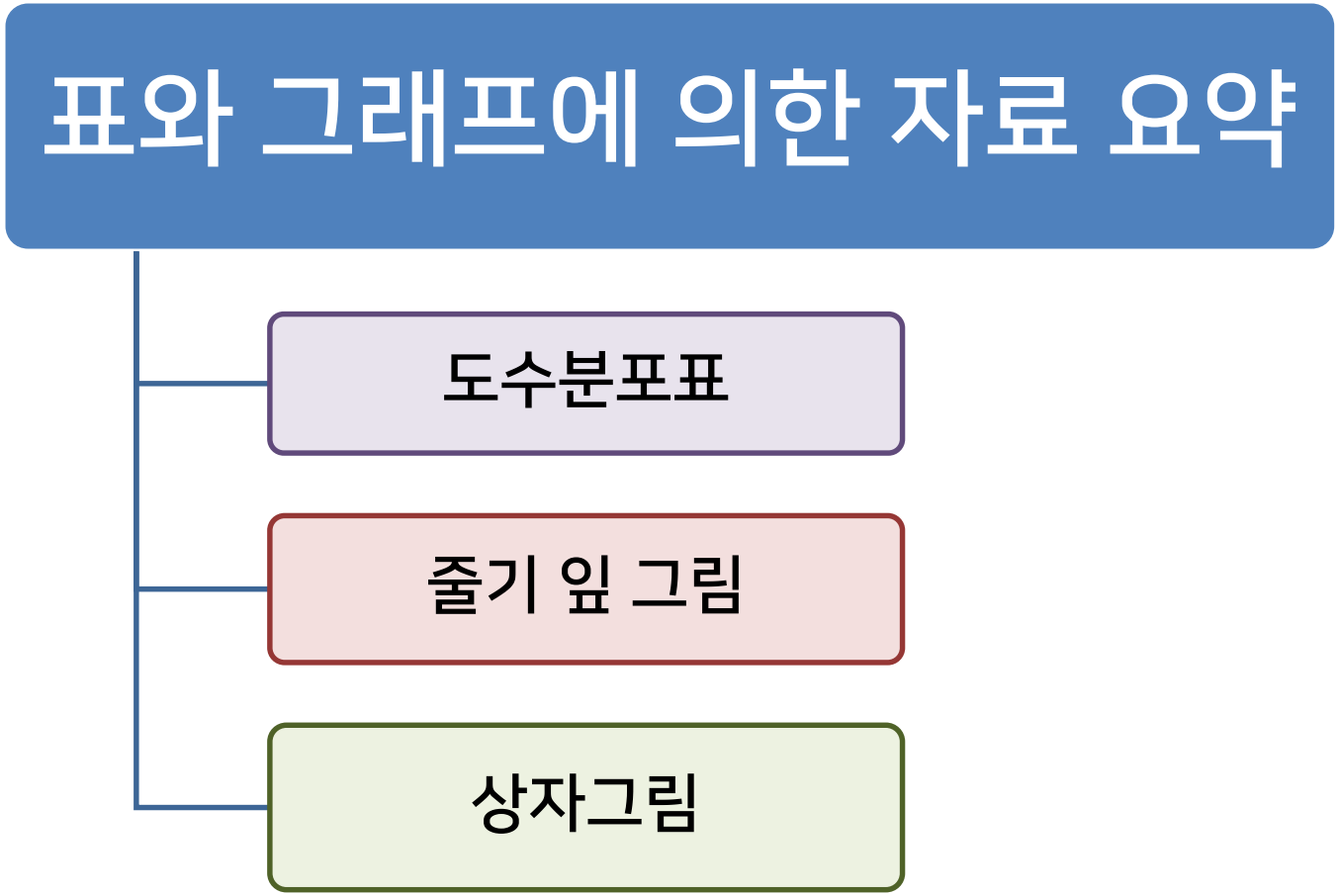


학습하기

3 4

4. 표와 그래프에 의한 자료요약

표와 그래프에 의한 자료 요약



도수분포표

도수분포표 [Frequency Table]

관측치들을 일정한 구간으로 나누어 각 구간에 속한 자료의 수를 정리한 표

작성 예

: 직장암 환자의 연령과 병의 진행 단계 자료

번호	연령	단계	번호	연령	단계
11	68	3	21	71	2
12	45	2	22	83	3
14	58	2	24	70	2
15	54	3	25	83	2
16	52	2	26	88	3
17	60	3	27	73	3

연속형 변수에 의한 도수분포표 →

계급구간	도수	상대도수	누적상대도수
40~49	2	0.050	0.051
50~59	9	0.225	0.275
60~69	7	0.175	0.450
70~79	13	0.325	0.775
80~89	9	0.225	1.000
계	40	1.000	1.000

도수분포표(Frequency Table)

도수분포표 [Frequency Table]

관측치들을 일정한 구간으로 나누어 각 구간에 속한 자료의 수를 정리한 표

작성 예

: 직장암 환자의 연령과 병의 진행 단계 자료

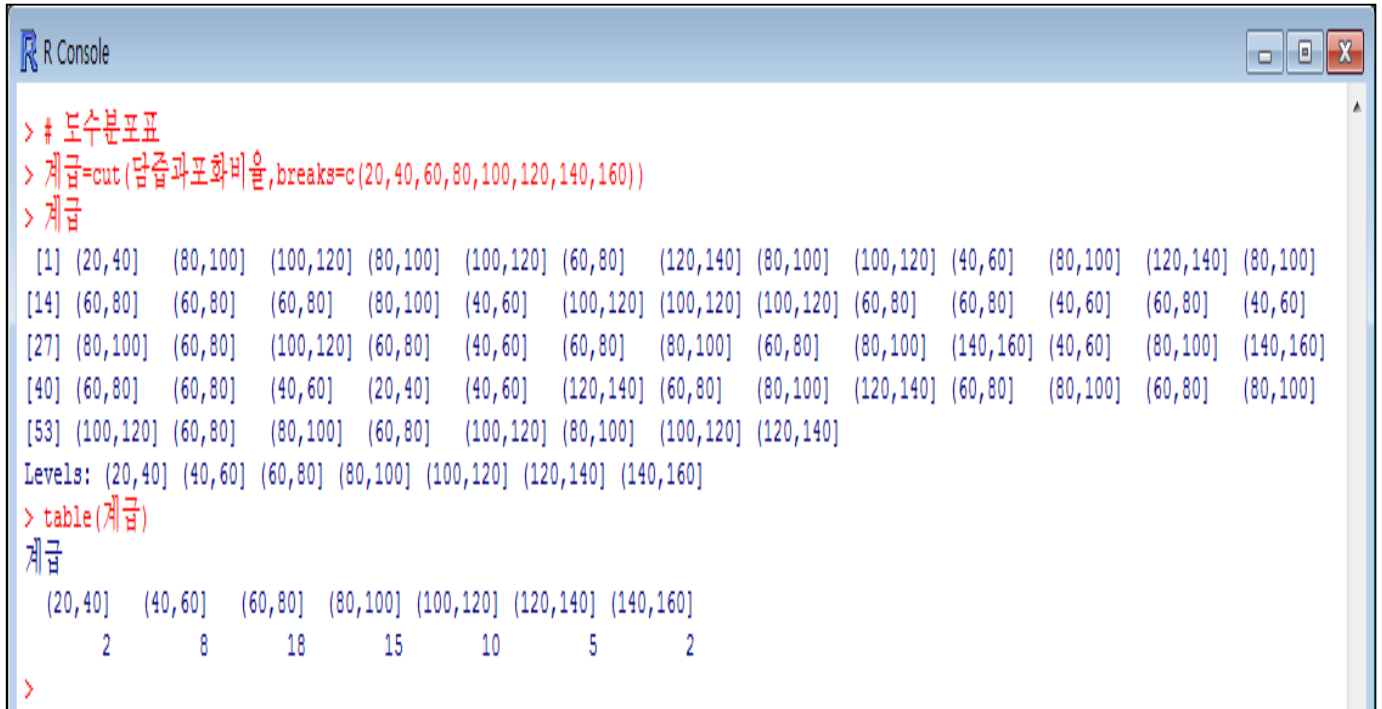
번호	연령	단계	번호	연령	단계
11	68	3	21	71	2
12	45	2	22	83	3
14	58	2	24	70	2
15	54	3	25	83	2
16	52	2	26	88	3
17	60	3	27	73	3

범주형 변수에 의한 도수분포표 ➡

단계	도수	상대도수	누적 상대도수
1	12	0.3	0.3
2	8	0.2	0.5
3	20	0.5	1.0
계	40	1.0	1.0

도수분포표


```
# 도수분포표
>계급=cut(담즙과포화비율,
        breaks=c(20,40,60,80,100,120,140,160))
>계급
>table(계급)
```



도수분포표 작성방법

작성방법

1. 관측치 중 최대값과 최소값을 찾는다.

번호	연령	단계	번호	연령	단계
11	68	3	21	71	2
12	45	2	22	83	3
14	58	2	24	70	2
15	54	3	25	83	2
16	52	2	26	88	3
17	60	3	27	73	3

2. 최대값과 최소값의 차이, 즉 범위를 구한다.

$\rightarrow$ $R = X_{man} - X_{min} = 88 - 45 = 43$

3. 몇 개의 구간으로 나눌 것인지 결정한다.

도수분포표 작성방법

4. 구간이 중복되지 않도록 경계값을 구한다

계급구간	도수	상대도수	누적 상대도수
40~49			
50~59			
60~69			
70~79			
80~89			
계			

5. 각 구간에 속한 관측치의 수를 세어 도수를 구한다.

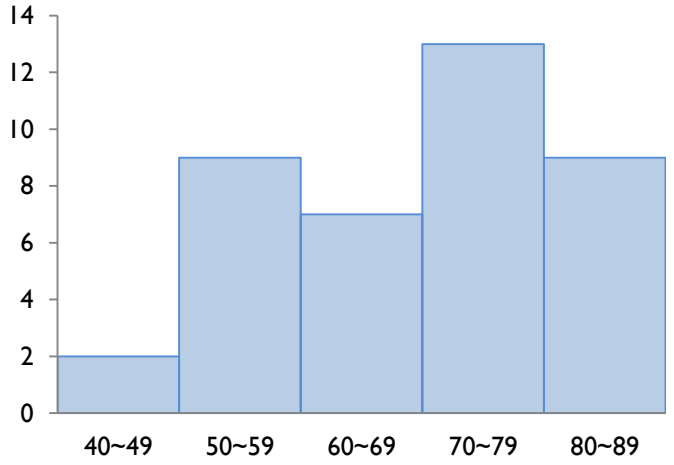
계급구간	도수	상대도수	누적 상대도수
40~49	2	0.050	0.051
50~59	9	0.225	0.275
60~69	7	0.175	0.450
70~79	13	0.325	0.775
80~89	9	0.225	1.000
계	40	1.000	1.000

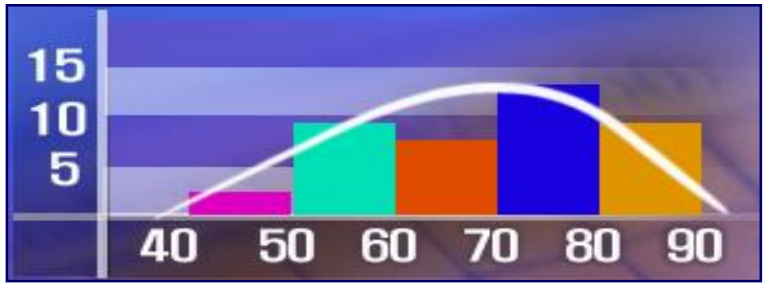
히스토그램(histogram)

히스토그램 [Histogram]

도수분포표를 가로축에 구간을, 세로축에 도수를 막대의 길이로 나타낸 그림

계급구간	도수	상대도수	누적 상대도수
40~49	2	0.050	0.051
50~59	9	0.225	0.275
60~69	7	0.175	0.450
70~79	13	0.325	0.775
80~89	9	0.225	1.000
계	40	1.000	1.000

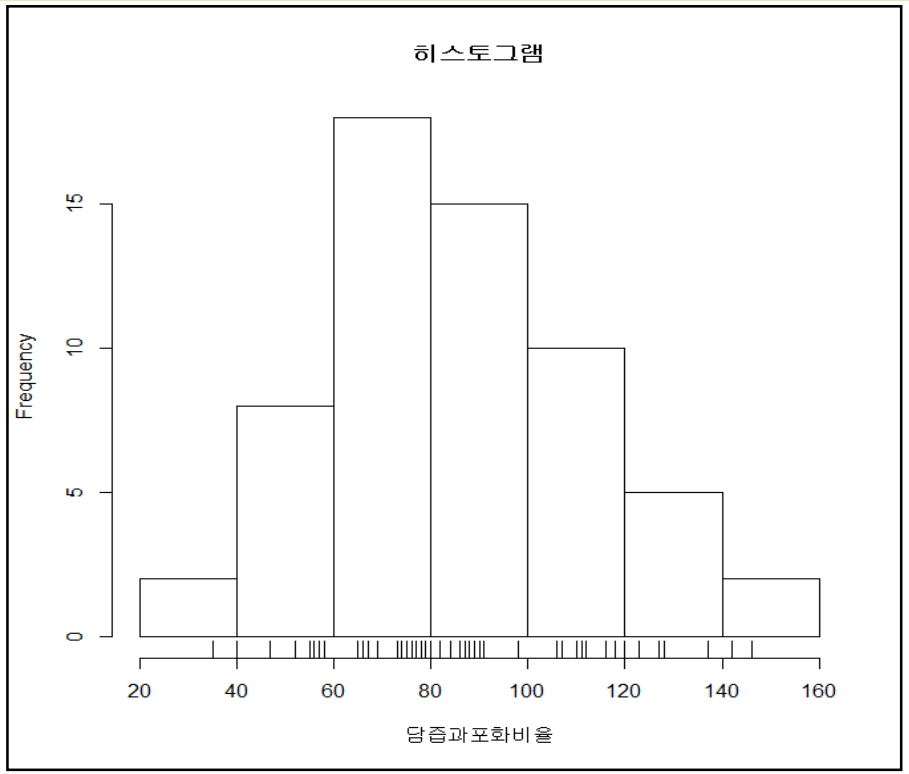




⇒자료의 분포 개형을 알 수 있다

히스토그램(histogram)


```
# 히스토그램
>hist(담즙과포화비율),
breaks=c(20,40,60,80,100,120,140,160),main="히스토그램")
>rug(담즙과포화비율)
```



줄기 잎 그림(stem and leaf Plot)

줄기 잎 그림 [Stem-and-leaf plot]

- 수치로 된 자료를 줄기와 잎으로 분류하여 자료의 분포개형을 파악하는 그림

- 54 → 50+4　　78 → 70+8

5	4
6	
7	8

줄기그림


```
# 줄기그림
>stem(담즙과포화비율)
# 줄기그림-줄기의 마디 2배로 늘이기
>stem(담즙과포화비율,scale=2)
```

R Console

```
>
> # 줄기그림
> stem(담즙과포화비율)

  The decimal point is 1 digit(s) to the right of the |

   2 | 5
   4 | 072256788
   6 | 556679334566789
   8 | 000244666778889018
  10 | 667001268
  12 | 033787
  14 | 26

>
>
```

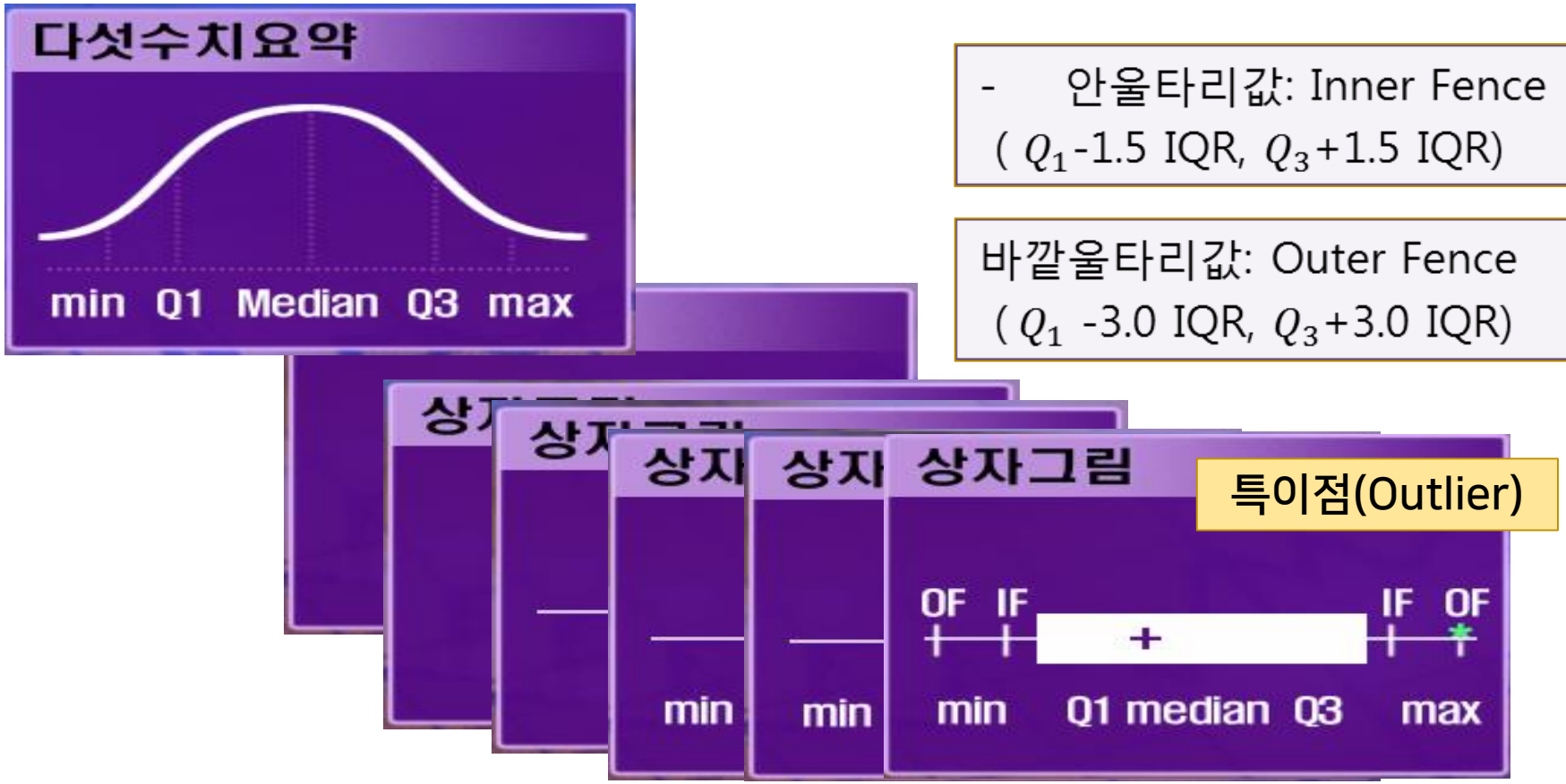
R Console

```
>
>
>
> # 줄기그림-줄기의 마디 2배로 늘이기
> stem(담즙과포화비율,scale=2)

  The decimal point is 1 digit(s) to the right of the |

   3 | 5
   4 | 07
   5 | 2256788
   6 | 556679
   7 | 334566789
   8 | 000244666778889
   9 | 018
  10 | 667
  11 | 001268
  12 | 03378
  13 | 7
  14 | 26
```

상자그림(Boxplot)

다섯수치 요약

[최소값, 제 1사분위수, 중위수, 제 3사분위수, 최대값]을 시각화한 그림



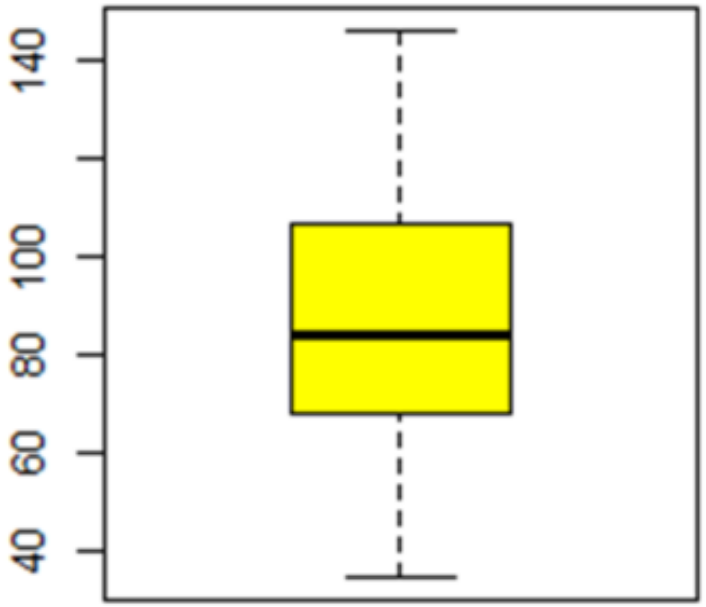
- 안울타리값: Inner Fence
(Q_1-1.5 IQR, Q_3+1.5 IQR)

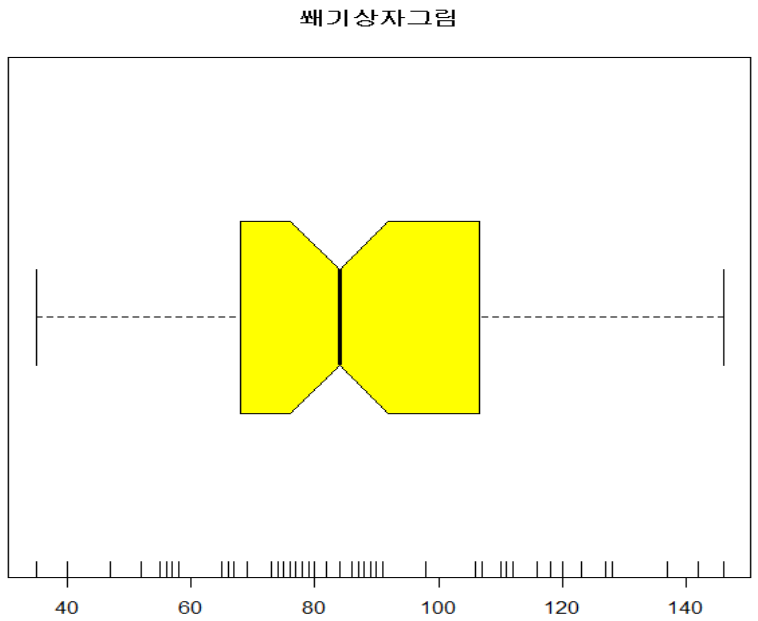
바깥울타리값: Outer Fence
(Q_1 -3.0 IQR, Q_3+3.0 IQR)

특이점(Outlier)

상자그림(Boxplot)

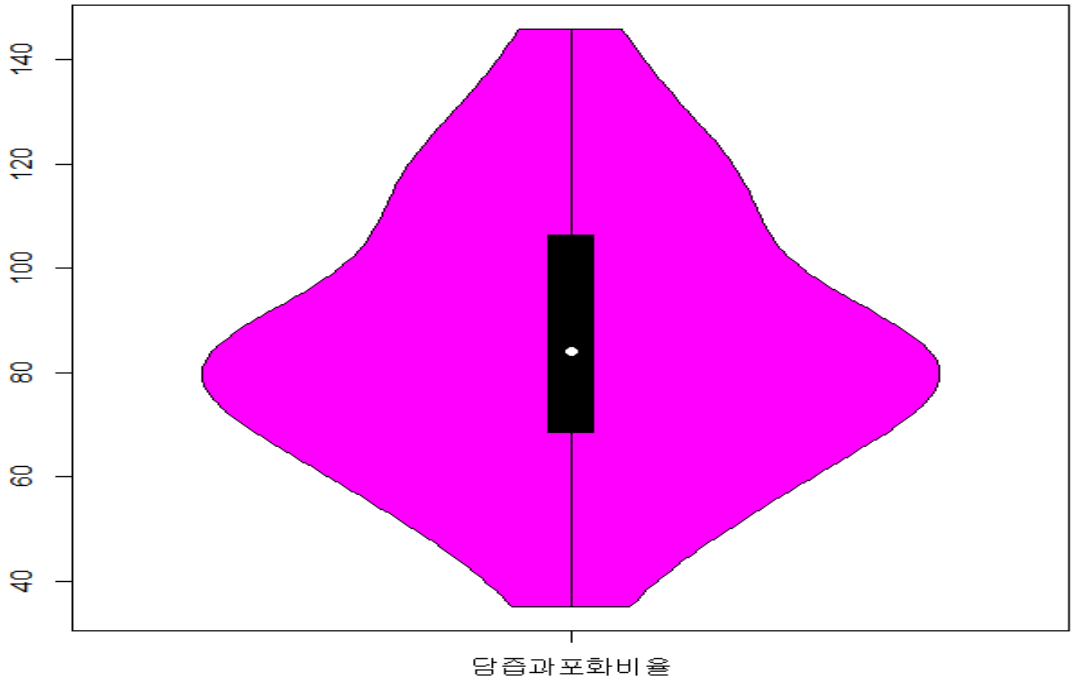

```
# 상자그림
>boxplot(담즙과포화비율,col="yellow",main="상자그림")
# 쐐기(notch)가 있는 상자그림(자료 병기)
>boxplot(담즙과포화비율,col="yellow",notch=T,horizontal=T,main="쐐기상자그림")
>rug(담즙과포화비율)
```





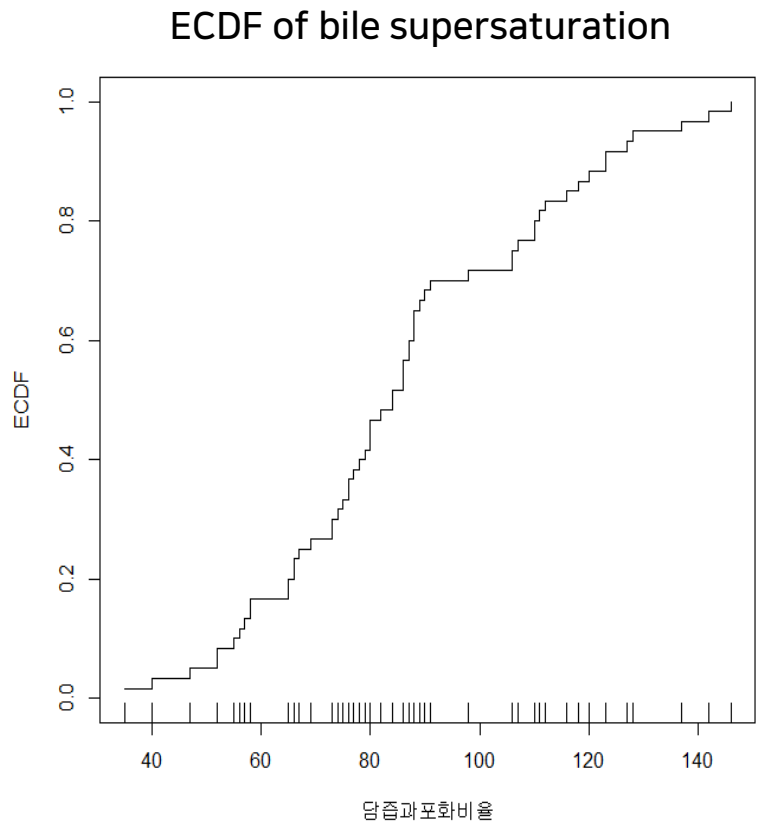
바이올린그림(히스토그램+상자그림 시너지효과)


```
# 바이올린그림(히스토그램+상자그림 시너지효과)
library(vioplot)
vioplot(담즙과포화비율,names="담즙과포화비율")
```



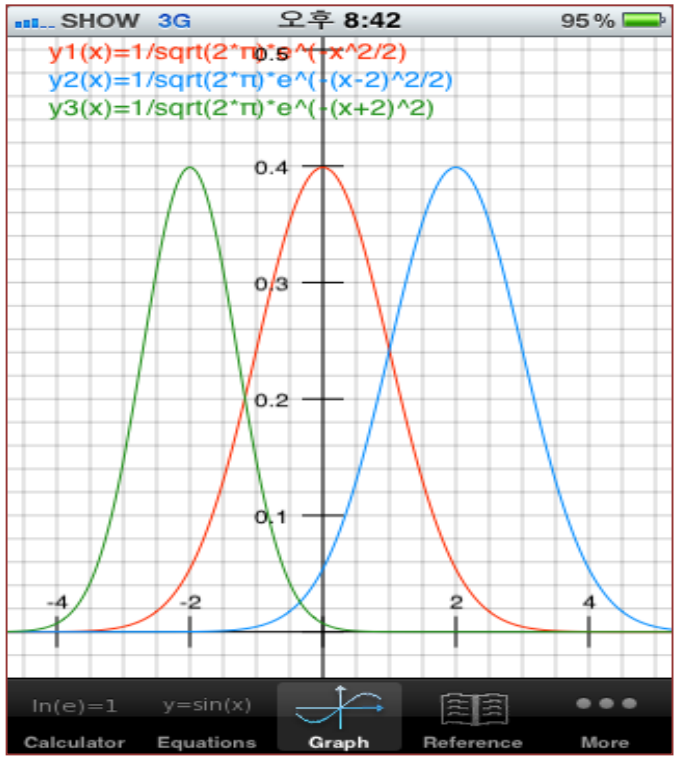
Ogive 경험누적분포함수

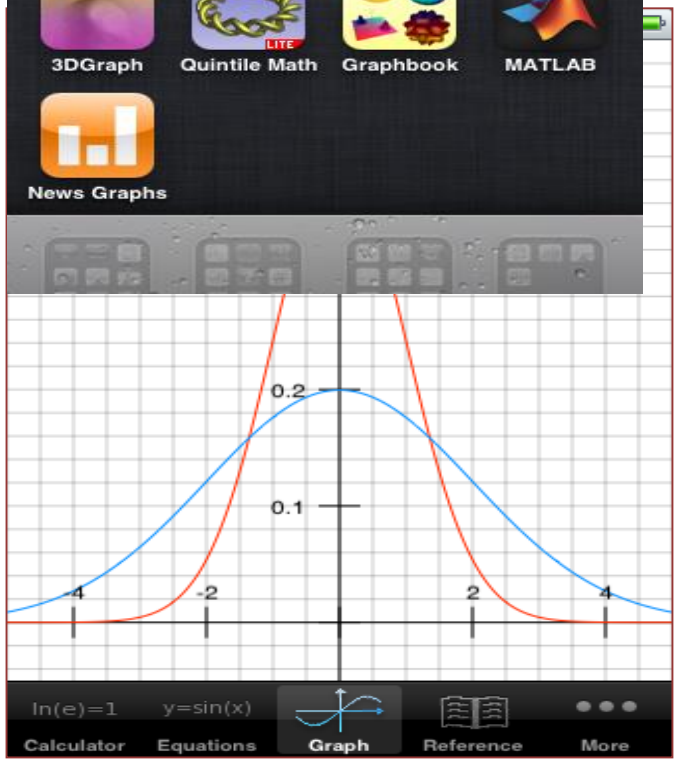

```
>plot(sort(담즙과포화비율),(1:n)/n,type="s",ylim=c(0,1),
main="ECDF of bile supersaturation",ylab="ECDF",xlab="담즙과포화비율")
>rug(담즙과포화비율)
```



그래프 작성 관련 무료앱

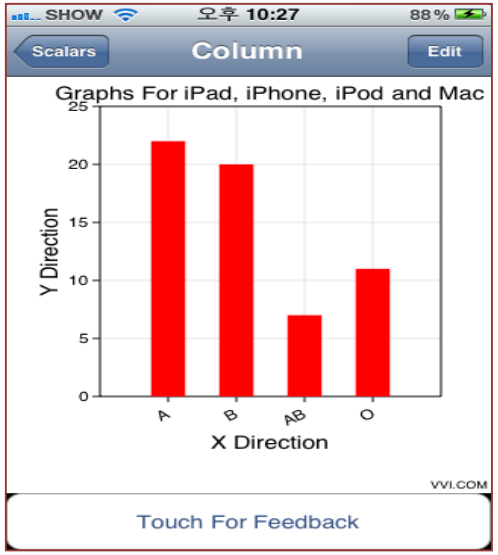
Mobile Applet for Graph

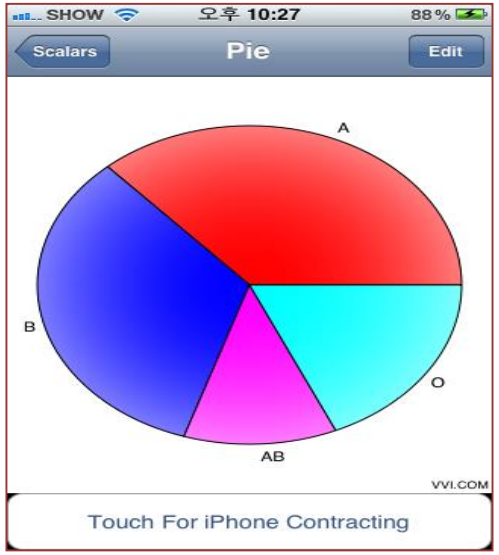


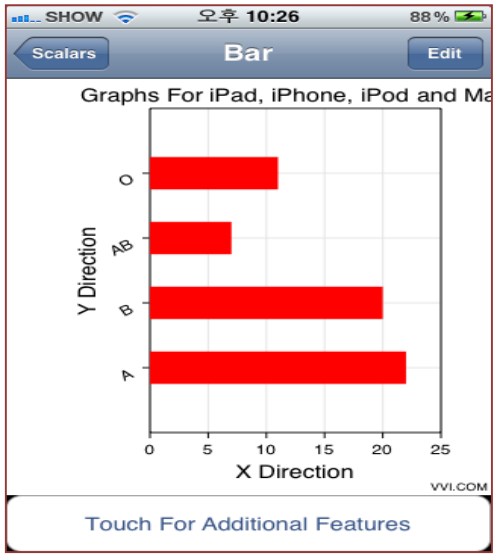


그래프 작성 관련 무료앱

Mobile Applet for Graph

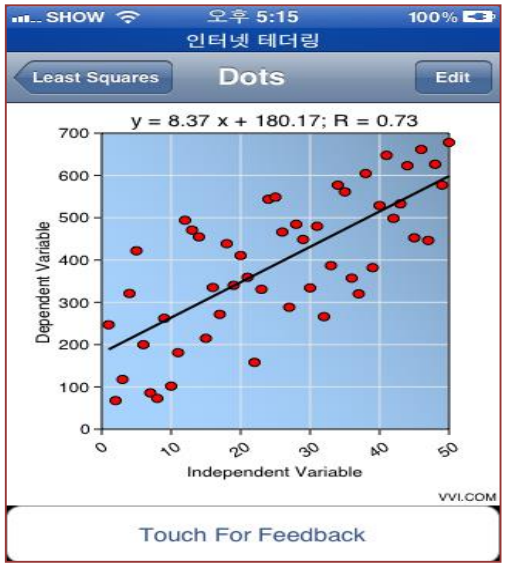


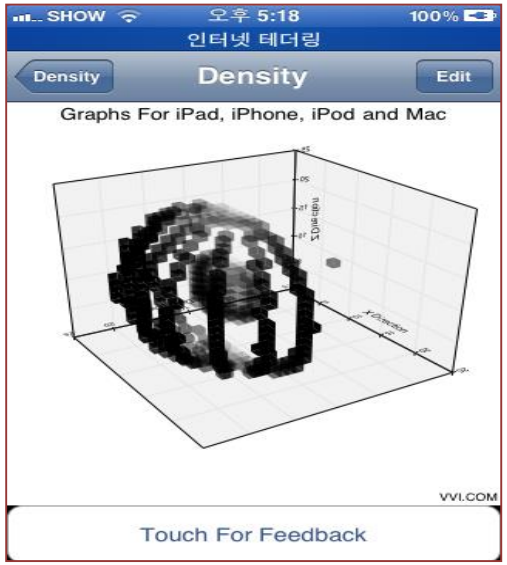


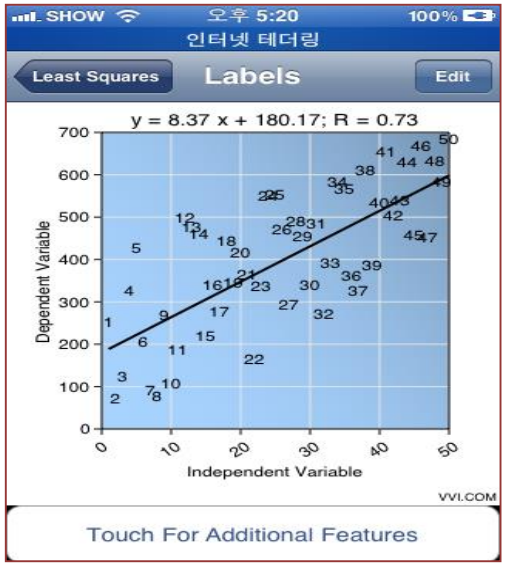


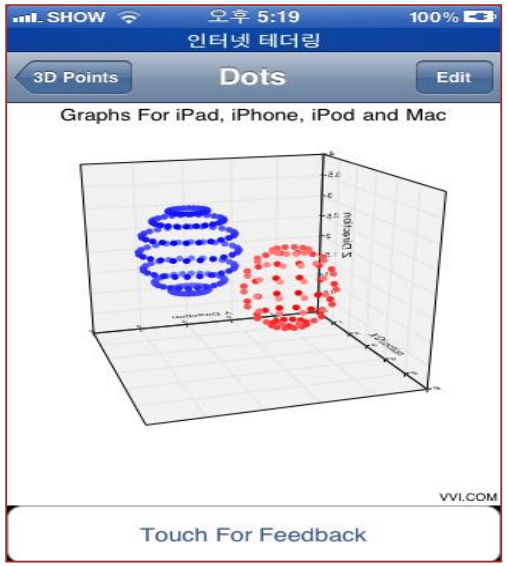
그래프 작성 관련 무료앱

Mobile Applet for Graph





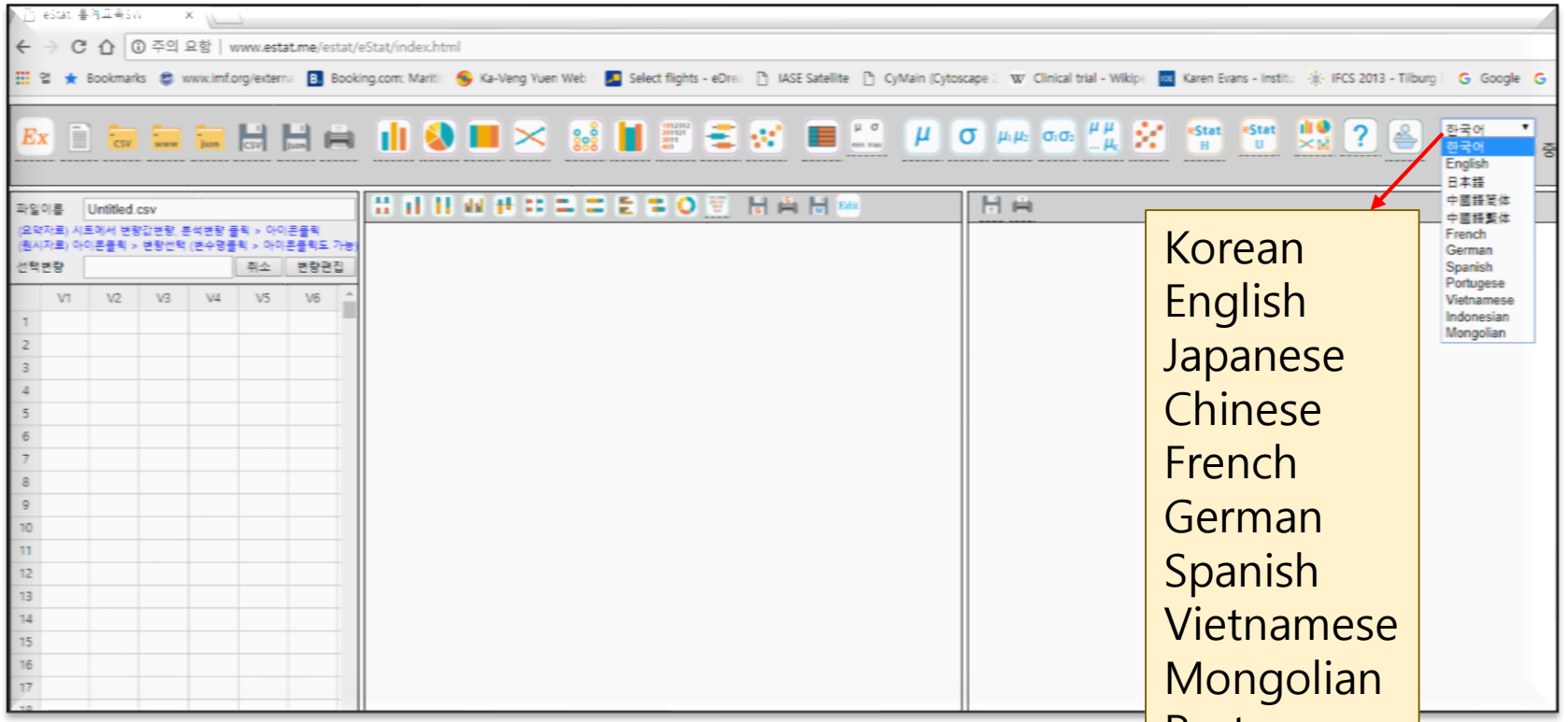




통계분석 무료 소프트웨어

eStat.me

http://www.estat.me/estat/eStatU/index.html

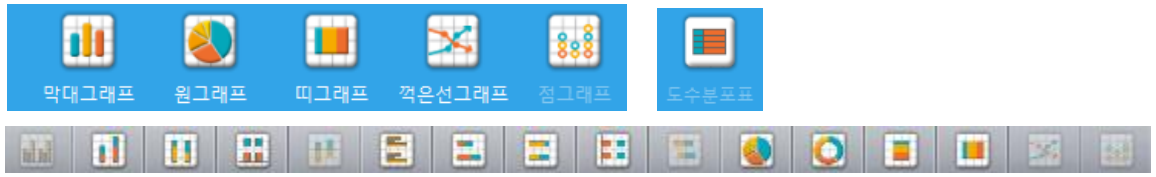


Korean
English
Japanese
Chinese
French
German
Spanish
Vietnamese
Mongolian
Portuguese
Polish

2. estat Summary

◎ Menu of Icon according to student level

- **Elementary school level**



- **Middle school level**

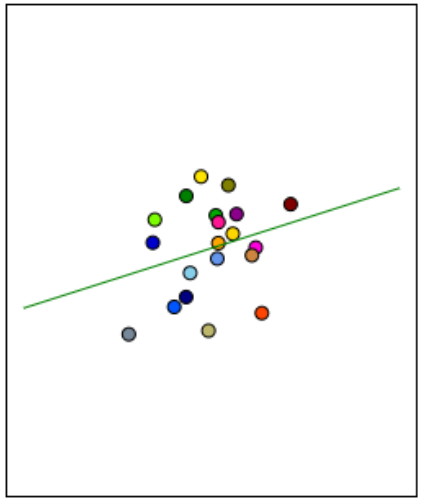
- **High school level**

- **University level**

2. estat Summary

eStatU - University Statistics Education SW

- Binomial Experiment
- Binomial Distribution
- Poisson Distribution
- Geometric Distribution
- HyperGeometric Distribution
- Exponential Distribution
- Normal Experiment
- Normal Distribution
- t Distribution
- ChiSquare Distribution
- F Distribution
- Law of Large Number
- Population vs Sample
- Dist of Sample Means
- Confidence Interval



Contact: jjlee@ssu.ac.kr

- Correlation Coeff
- Regression Experiment

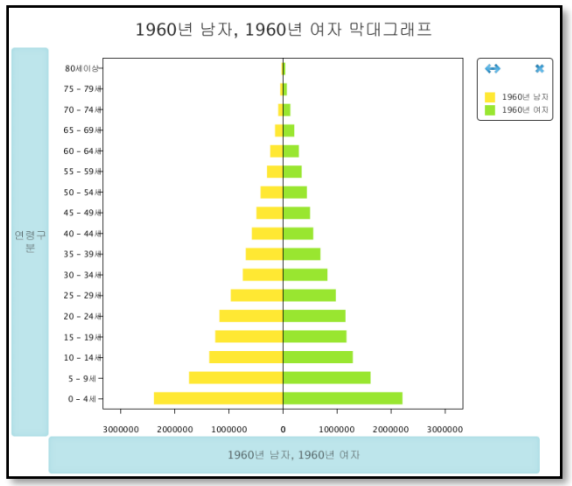
- Testing Hypothesis μ
- Testing Hypothesis σ^2
- Testing Hypothesis P
- Testing μ with α, β
- Testing Hypothesis μ_1, μ_2
- Testing Hypothesis σ_1^2, σ_2^2
- Testing Hypothesis P_1, P_2
- Testing Hypothesis ANOVA
- Sign Test
- Wilcoxon Signed Rank Sum Test
- Wilcoxon Rank Sum Test
- Kruskal-Wallis Test
- Friedman Test
- Goodness of Fit Test
- Testing Independence

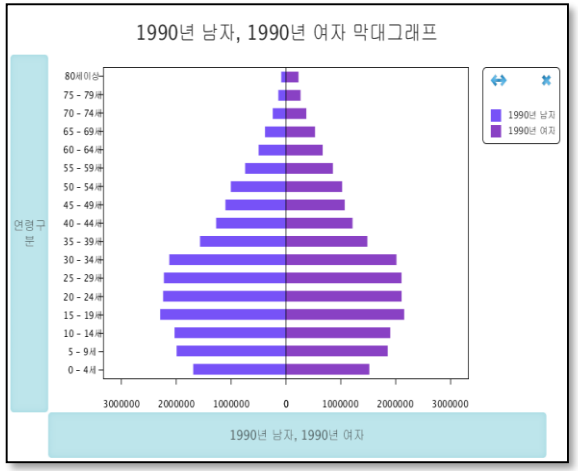
2. estat Summary

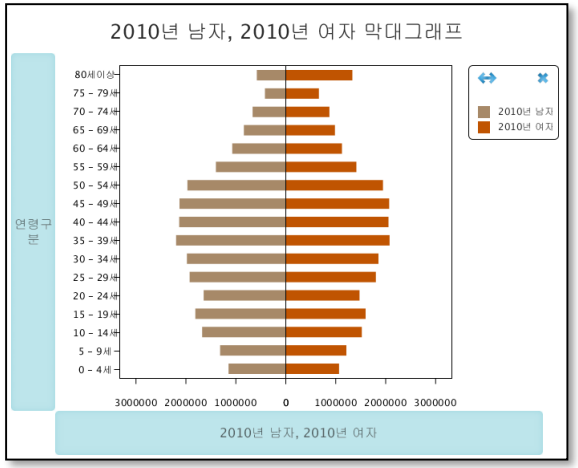
◎ Masking function of one record dot, variable,

2. estat Summary

◎ 2 Direction Bar chart(Census Pyramid)

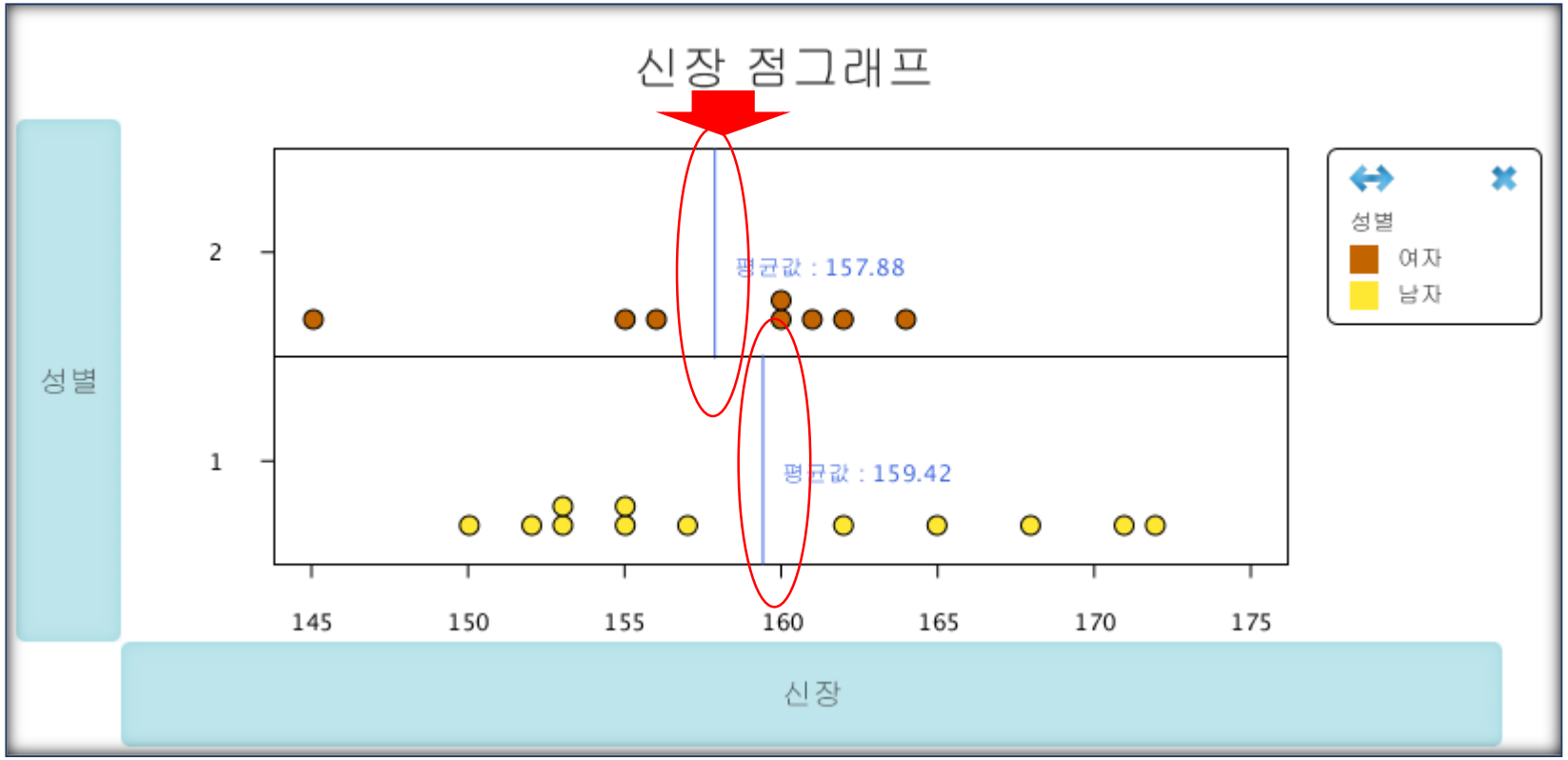






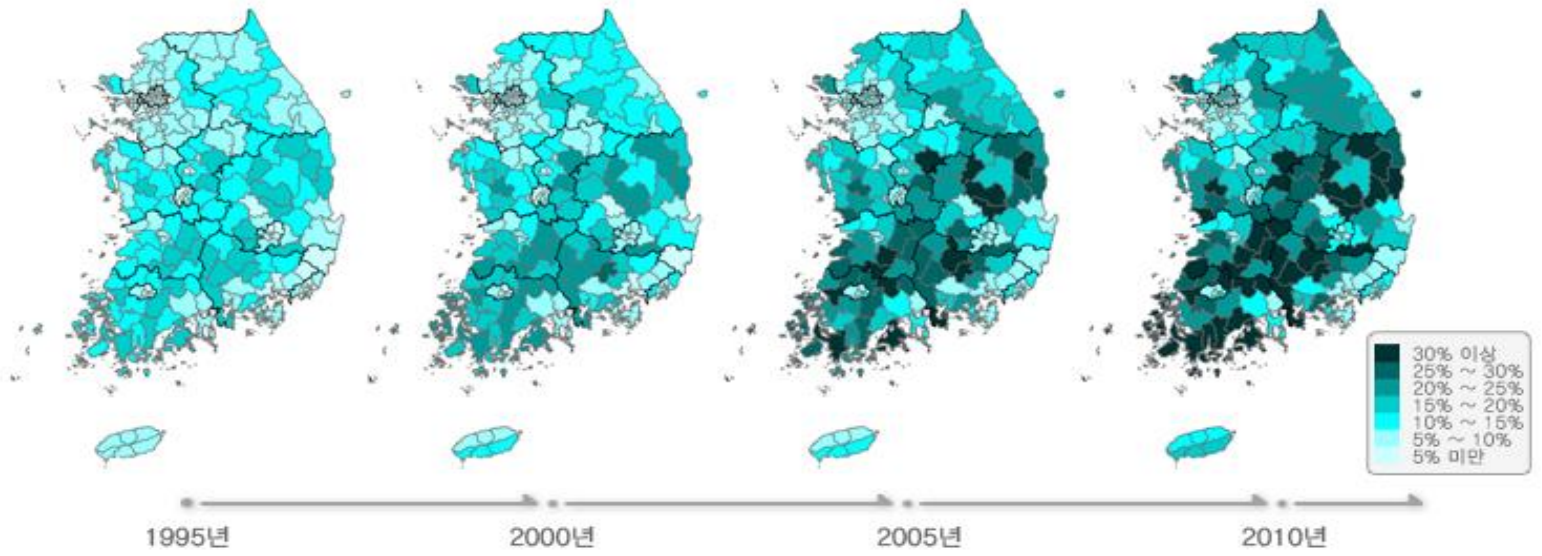
2. estat Summary

◎ Dot diagarm & mean line for hight comparison



2. estat Graph

◎ Geographic graph(GIS)



학습하기

4 5

5. 추정과 검정

추정과 가설검정

추정과 가설검정

- 점추정
- 구간추정
- 가설검정의 개념

추정과 검정

통계적 추론 [statistical inference]

- 모집단의 일부인 표본을 추출하여 모집단의 특성을 알아보는 것

점추정 [point estimation]

- 모수를 하나의 수치로 추정하는 것

[평균, 비율, 분산, 평균]

구간추정 [interval estimation]

- 모수가 속할 구간을 두 수치로 추정하는 것

폐경 연령 구간 (θ_L, θ_u) = (54.1, 56.2)

가설검정

가설검정 [Hypothesis Testing]

모수에 대한 가설의 옳고 그름을 판단하는 절차

단계1 가설을 세운다

귀무가설 [null hypothesis H_0]

- 대립가설에 반하는 가설로 차이가 없다는 것을 내용으로 한다.

대립가설 [alternative hypothesis H_1]

- 연구자의 주장을 담고 있는 가설

자료를 통하여 대립가설을 입증할 강력한 증거가 있을 때는 연구자의 주장[H_0]을 택하고 증거가 충분치 못하면 귀무가설 [H_1]로 돌아간다.

가설검정

가설검정 [Hypothesis Testing]

제 1종 오류 [type 1 error]

- α - 귀무가설이 사실일 때 우연에 의해 귀무가설을 기각하게 되는 오류

유의수준 [significance level]

- 제 1종 오류가 일어날 확률의 최대 허용치

제 2종 오류

- β - 대립가설이 사실일 때 거짓인 귀무가설을 기각하지 않아 발생 하는 오류

검정력 [power of test]

- 귀무가설이 사실이 아닐 때 귀무가설을 기각하게 될 확률

가설검정 [Hypothesis Testing]

- 유의확률[P-value]
  - 귀무가설을 기각할 만한 충분한 증거가 되는 기준이 되는 확률값
  - 귀무가설 하에서 관측된 사건 이상으로 귀무가설에 반하는 사건이 일어날 확률로 P값이 작을수록 대립가설에 대한 강한 증거가 된다
  - 통계적으로 유의(significant)하다고 해석

가설검정

가설검정 [Hypothesis Testing]

- 단측검정 [one-sided test]
  - 주장하는 바가 어느 한 방향으로 이루어지는 검정
  - 가설형태 $H_0 : P_1 \leq P_0$

    $H_1 : P_1 \leq P_0$
- 양측검정 [two-sided test]
  - 비교하고자 하는 두 모수의 차이가 있는가 만을 검정하는 가설
  - 가설형태 $H_0 : P_1 = P_0$

    $H_0 : P_1 \neq P_0$

가설검정의 절차

1. 귀무가설과 대립가설을 세운다 $[H_0, H_1]$
2. 유의수준을 α로 정한다.
3. 관심 통계량을 정한다.
4. 검정 통계량을 계산한다.
5. 유의확률[P-value]에 의해 결론을 내린다.

다음시간에는

- 2강 보건정보 데이터의 기초분석
- 3강 연속형자료의 분석
- 4강 범주형자료 분석