탐색적 자료분석

서울대학교 통계연구소

2024년 2월

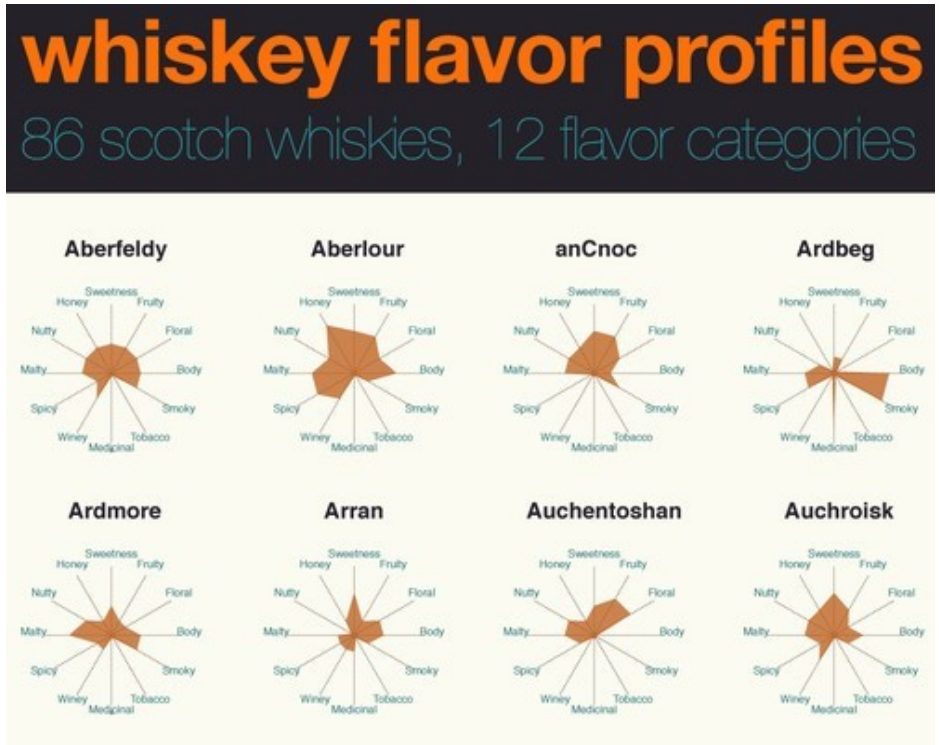
이번 강의에서 다룰 내용

- 정보(데이터)의 시각화
- R 그래픽스 -plot() 중심으로
- 기술통계를 통한 데이터 요약
- 통계적 추론 - 모평균 추정 및 검정

정보의 시각화

- Data explosion-빅데이터
  - 뉴욕 타임즈가 하루에 싣는 정보의 양은 17세기 영국의 평범한 한 사람이 평생 소비하는 정보의 양과 비슷 (Wurman, S.A. (1987) "Information Anxiety" New York: Doubleday)
  - 페이스북에서는 하루에 30페타 바이트 이상의 정보가 저장, 공유 (http://wikibon.org/blog/taming-big-data/)
  - 전 세계에서 하루에 50억 건 이상의 트윗, 문자 메시지 등이 사용 (http://www.sas.com/resources/whitepaper/wp_4634
  - 정보의 양이 증가할 수록 정보를 다루는 일이 더욱 어려워 진다 - 시각화의 중요성

인포그래픽 (infographics)



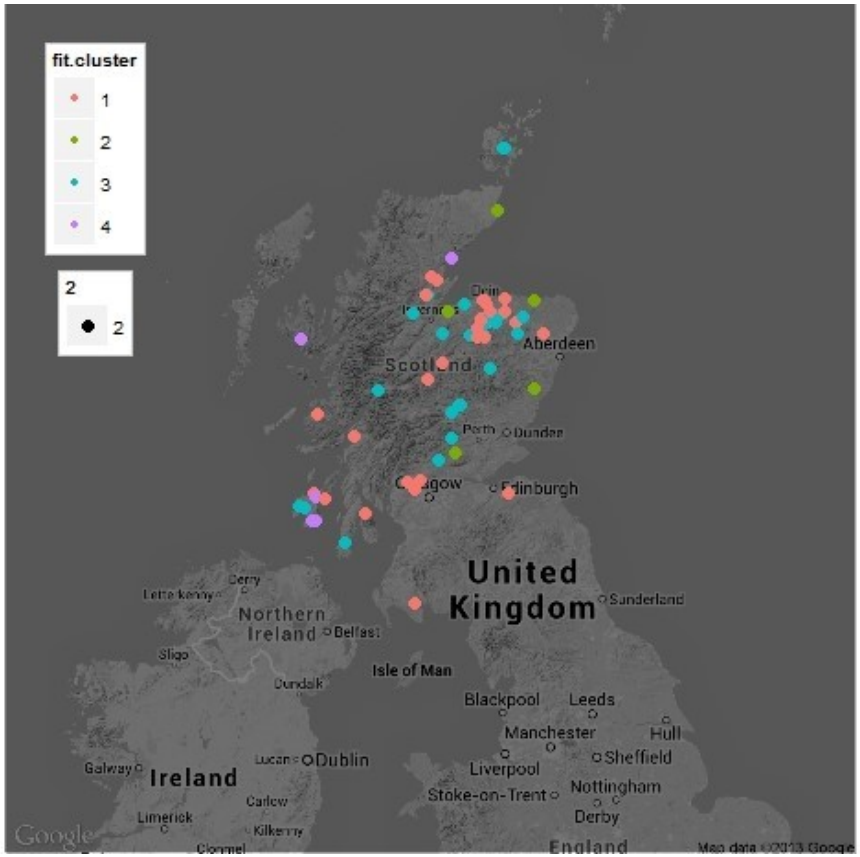
WonkViz Blog

http://wonkviz.tumblr.com/post/72159021235/whiskey-flavor-profiles

GIS 를 이용한 시각화

다른 정보와 mash-up

데이터 군집화(clustering) + 구글지도 매시업(mash-up)



– Revolution Analytics Blog

http://blog.revolutionanalytics.com/2013/12/k-means-clustering-86-single-malt-scotch-whiskies.html

텍스트 시각화

- 빅데이터 연구에서 텍스트 분석의 중요성이 점차 커짐
  - 최근 수년간 급속도로 성장한 소셜네트워크 서비스의 영향
- 텍스트는 숫자와 달리 명목 데이터 (nominal data) 이기 때문에 그래프로 표현하기 어려워 몇가지 전처리 과정을 거쳐 시각화
  - 태그클라우드 (tag cloud) - 단어의 노출 빈도수를 계산하여 시각화
  - 워드트리 (word tree) - 문장 내에서 단어 간 연결 구조를 시각화

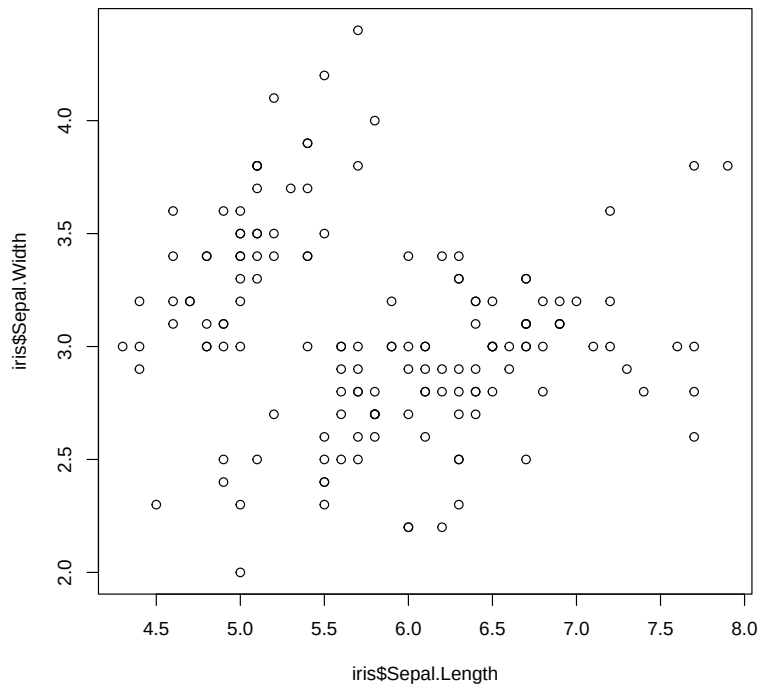
간단한 예시

```
library(tm)
library(wordcloud)
wordcloud(c(letters, LETTERS, 0:9),seq(1, 1000, len = 62))
```

R 그래픽스- plot()

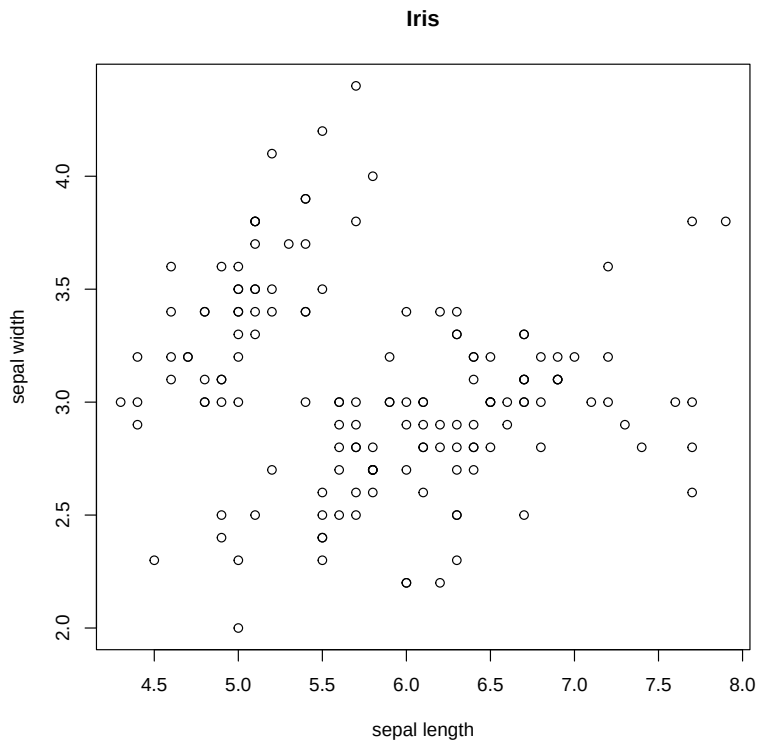
R에서 데이터를 그래프로 나타내는 데 가장 기본이 되는 함수

```
data(iris)
plot(iris$Sepal.Length,iris$Sepal.Width)
```



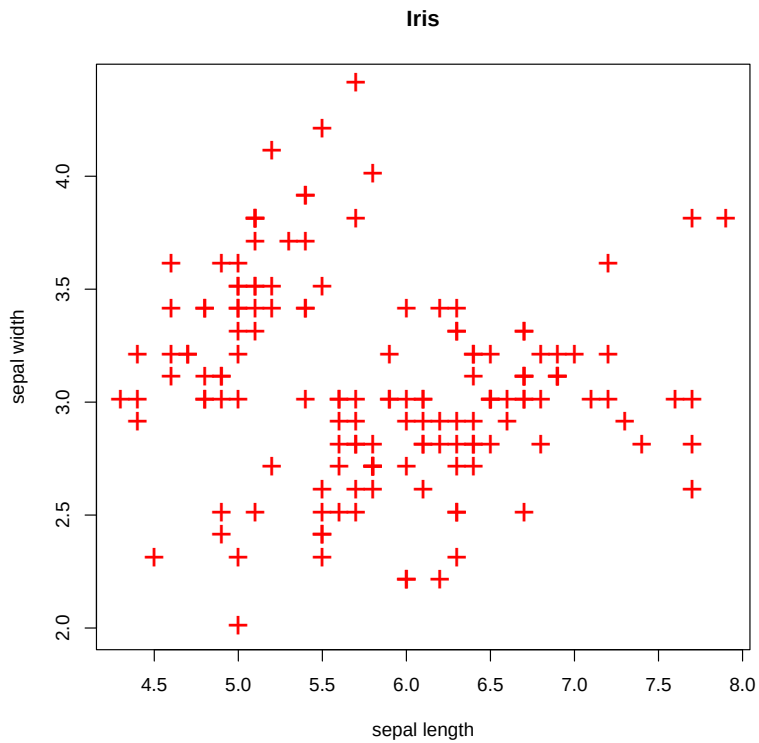
이름 붙이기: ‘xlab‘, ‘ylab‘, ’main’

```
plot(iris$Sepal.Length,iris$Sepal.Width, xlab="sepal length",
     ylab="sepal width",main='Iris')
```



점의 특성 바꾸기: pch, cex,col

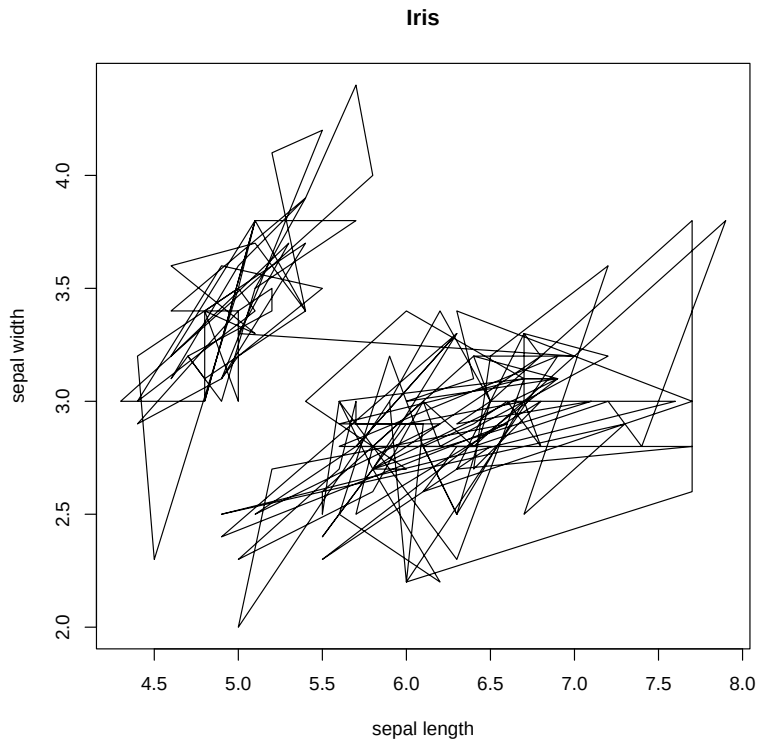
```
plot(iris$Sepal.Length,iris$Sepal.Width, xlab="sepal length",
     ylab="sepal width", main="Iris", pch='+',cex=2,col='red')
```



- pch 에 보여줄 수 있는 심볼의 목록은 ‘?points‘로 찾아볼 수 있다 (‘points()‘ 함수에 대해서는 뒤에 배운다).
- 사용 가능한 색깔의 이름들은 ‘colors()‘ 함수를 실행하여 볼 수 있다.

점들을 선으로 잇기: type

```
plot(iris$Sepal.Length,iris$Sepal.Width, xlab="sepal length",
     ylab="sepal width", main="Iris", type='l')
```

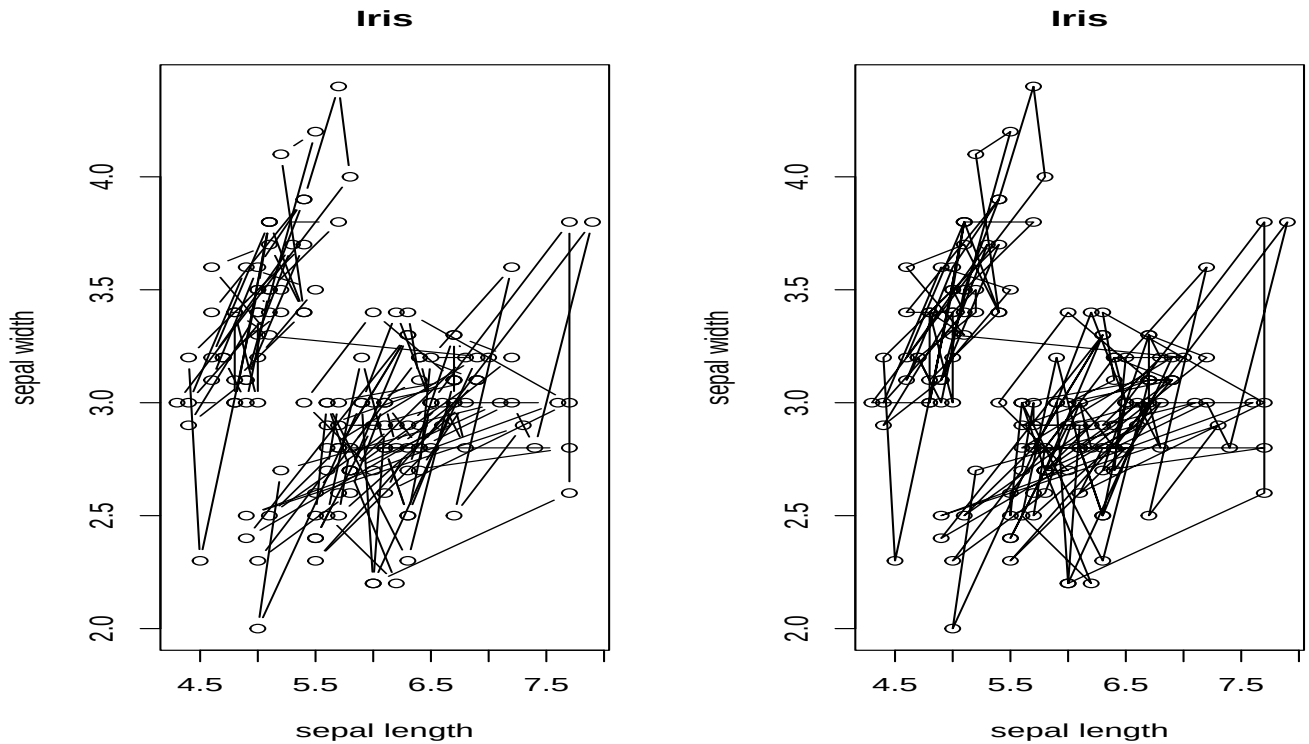


다른 옵션은 ‘?plot‘에서 ‘type‘ 섹션 참조.

한창에 여러개의 그래프 그리기-mfrow

par(mfrow = c(행 수, 열 수))

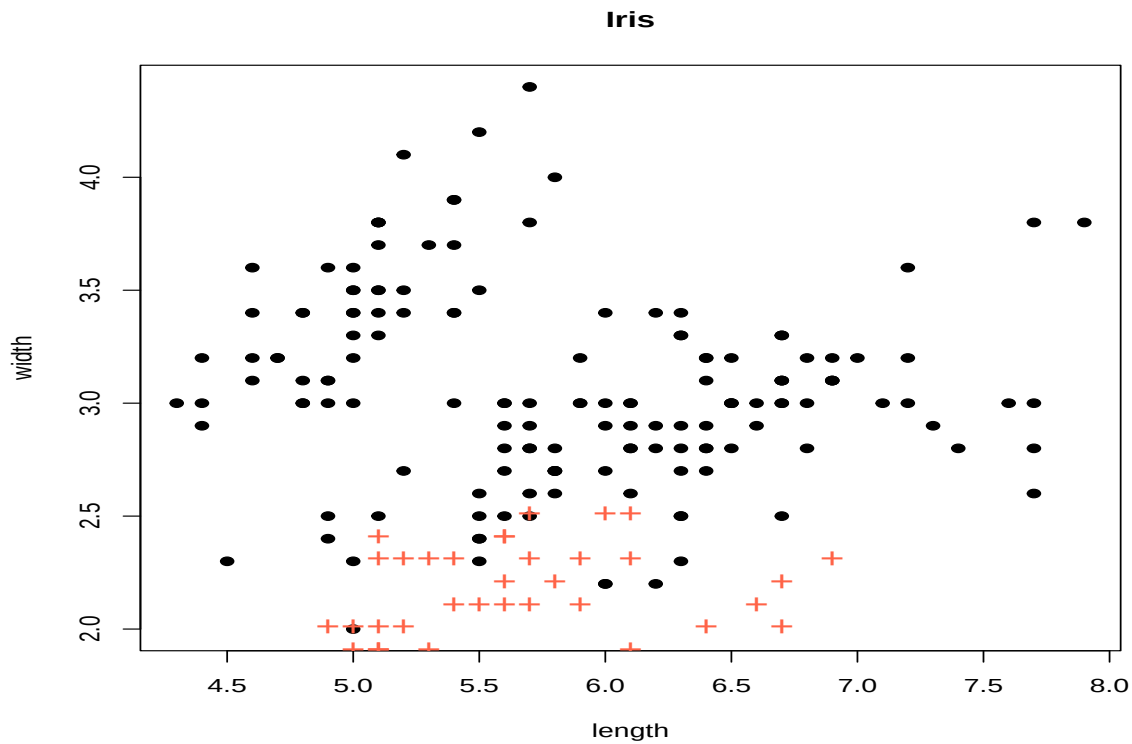
```
oldpar <- par(mfrow=c(1, 2))     # par문은 이전 설정을 반환함
plot(iris$Sepal.Length,iris$Sepal.Width, xlab="sepal length",
     ylab="sepal width", main="Iris", type='b')
plot(iris$Sepal.Length,iris$Sepal.Width, xlab="sepal length",
     ylab="sepal width", main="Iris", type='o')
par(oldpar)    # 이전 설정 회복
```



점: points()

이미 생성된 plot에 점을 추가로 그려준다.

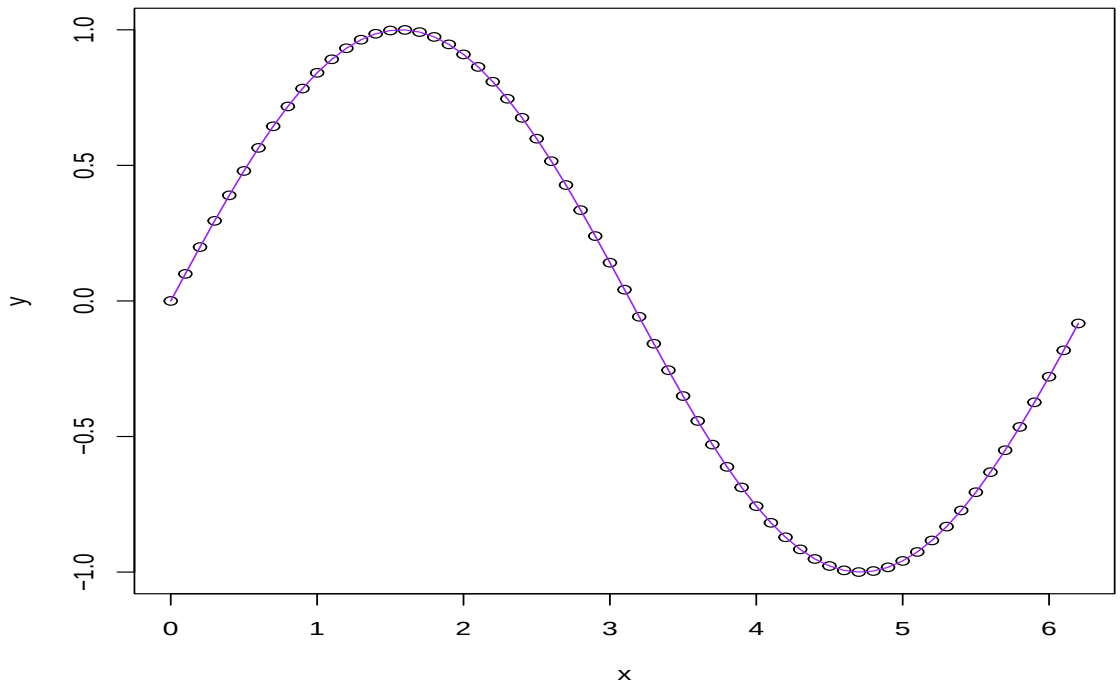
```
attach(iris)
plot(Sepal.Length , Sepal.Width , cex=1.5, pch=20, xlab="length",
     ylab="width", main="Iris")
points(Petal.Length , Petal.Width , cex=1.5, pch="+", col="tomato")
```



선: lines()

이미 생성된 plot에 선을 추가로 그려준다.

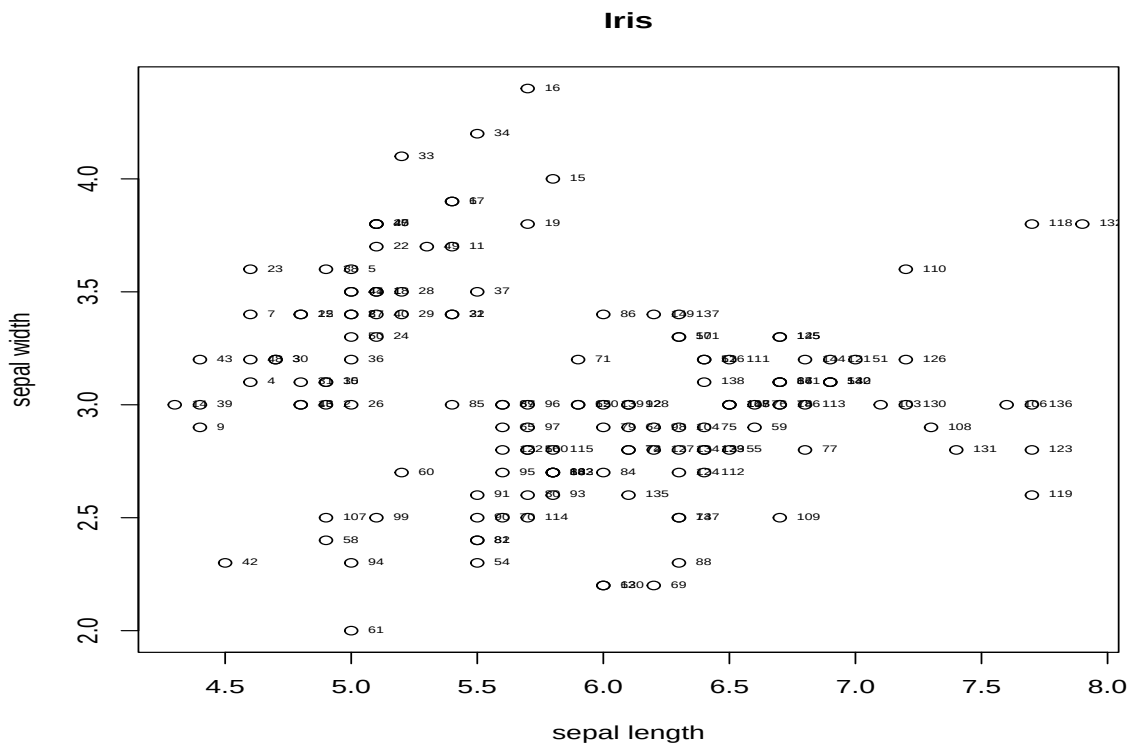
```r
x <- seq(0, 2*pi, 0.1)
y <- sin(x)
plot(x, y)
lines(x, y, col="purple")
```



문자추가: text()

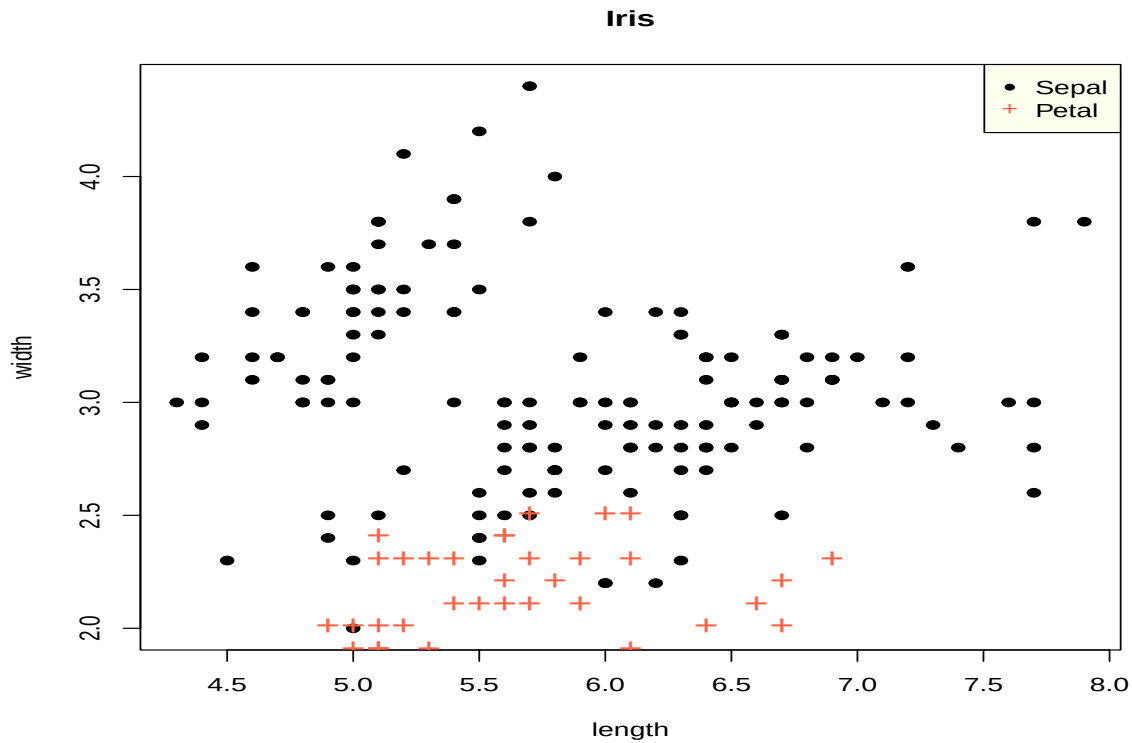
이미 생성된 plot에 문자를 표시-좌표의 오른쪽에 텍스트 표시

```
plot(iris$Sepal.Length,iris$Sepal.Width, xlab="sepal length",
     ylab="sepal width", main="Iris")
text(iris$Sepal.Length,iris$Sepal.Width, pos=4, cex=0.5 )
```



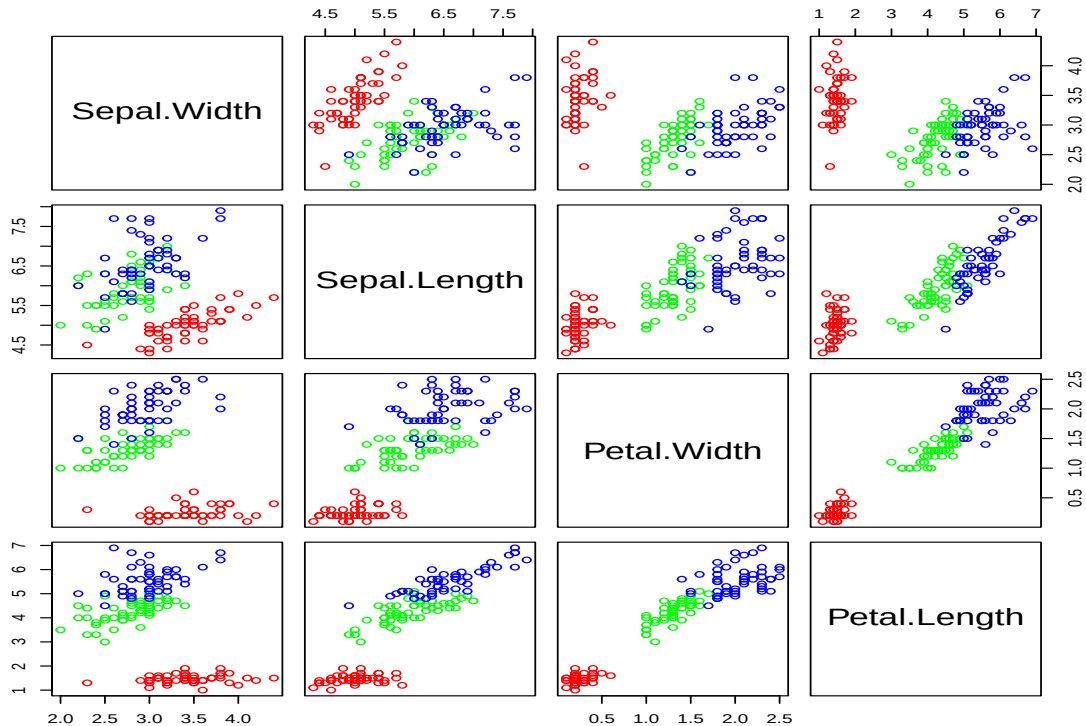
범례 : legend()

```
plot(iris$Sepal.Length , iris$Sepal.Width , cex=1.5, pch=20,
     xlab="length", ylab="width", main="Iris")
points(iris$Petal.Length , iris$Petal.Width , cex=1.5, pch="+",
       col="tomato")
legend("topright", legend=c("Sepal", "Petal"), pch=c(20, 43),
       col=c("black", "tomato"), bg="ivory")
```



다변수 데이터의 산점도: pairs()

```
pairs(~Sepal.Width + Sepal.Length + Petal.Width + Petal.Length,
      data=iris, col=c("red", "green", "blue")[iris$Species])
```



기술통계와 통계적 추론

- 기술 통계학(descriptive statistics)은 측정이나 실험에서 수집한 자료의 정리, 표현, 요약, 해석 등을 통해 자료의 특성을 규명하는 통계적 방법이다. ([위키피디아](http://ko.wikipedia.org/wiki/
  - 중심 경향성(central tendency)의 측도: 평균(mean), 중앙값(median), 최빈값(mode)
  - 산포도(dispersion)의 측도: 범위(range), 분산(variance), 표준편차(standard deviation)
- 통계적 추론(statistical inference)은 모집단에 대한 어떤 미지의 양상을 알기 위해 통계학을 이용하여 추측하는 과정을 지칭한다. ([위키피디아](http://ko.wikipedia.org/wiki/
- 표본에 대해서는 언제나 정리, 표현, 요약, 해석 등을 할 수 있으나, 이를 모집단 전체에 대해 일반화하는 것은 관측되지 않은 데이터가 항상 있기에 쉽지 않은 일이다.

자료요약- 자료의 종류

- 이산형 자료
  - 계수형 자료 (counting): 보험가입 건수, 차량 생산 대수
  - 범주형 자료(categorical): 성별, 인종별, 나이대별 (명목형, 순서형)
- 연속형 자료: 자료의 값이 연속적인 어떤 구간에서 관측되는 경우
  - 예시: 기온, 전등의 지속시간

- 자료의 형태에 따라 특징을 알아내기 위해 다양한 종류의 그래프로 표현해 본다.
- 이산형: 도수 분포표, 막대그래프, 원그래프
- 연속형: 히스토 그램, 상자그림

이산형자료에서의 요약

도수 분포표 (frequency table)

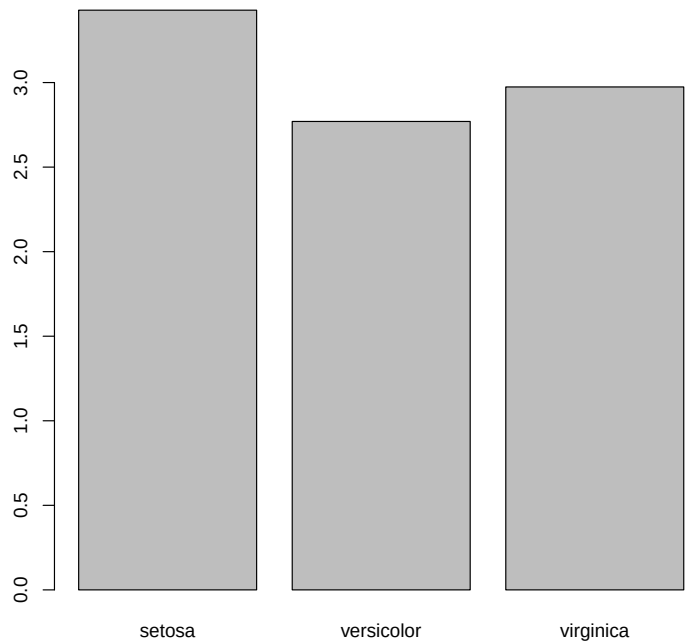
- 자료가 가질수 있는 값을 어떤 특성에 의하여 구분하여 해당 특성을 갖는 자료의 개수(도수, frequency)를 표로 나타낸 것.
- 주사위를 120번 던졌을때,

주사위 눈	1	2	3	4	5	6	합계
도수	18	25	14	22	26	15	120
상대도수	0.15	0.21	0.12	0.18	0.22	0.12	1

막대그래프

- 수평축에 특성값을 놓고 막대의 높이가 도수나 상대도수에 비례하도록 그린다.

```
barplot(tapply(iris$Sepal.Width,iris$Species,mean))
```



원그래프

- 중심각의 크기나 넓이가 상대도수에 비래하도록 그림

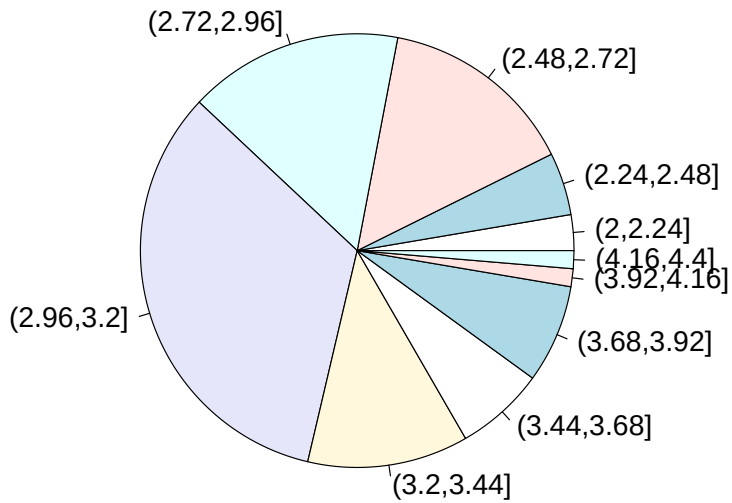
```
# 데이터의 범위를 10개 구간으로 나눔
partition <- cut(iris$Sepal.Width , breaks=10)
head(partition,3)

[1] (3.44,3.68] (2.96,3.2]  (2.96,3.2]
10 Levels: (2,2.24] (2.24,2.48] ... (4.16,4.4]

table(partition)     # 구간별 데이터 개수 세기

partition
   (2,2.24] (2.24,2.48] (2.48,2.72] (2.72,2.96]
          4           7          22          24
 (2.96,3.2]  (3.2,3.44] (3.44,3.68] (3.68,3.92]
         50          18          10          11
(3.92,4.16]  (4.16,4.4]
          2           2
```

```r
pie(table(cut(iris$Sepal.Width , breaks=10)), cex=1.5)
```

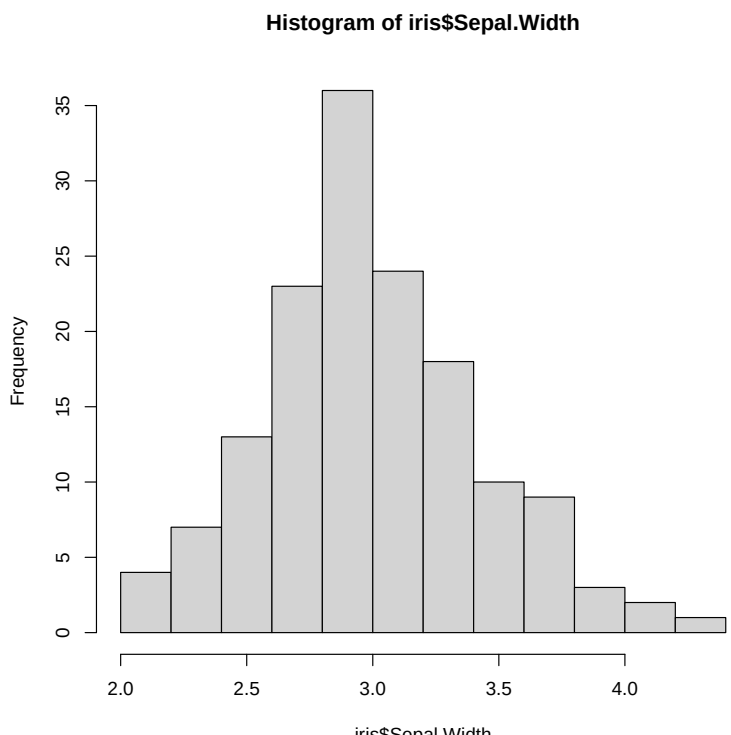


연속형자료에서의 요약

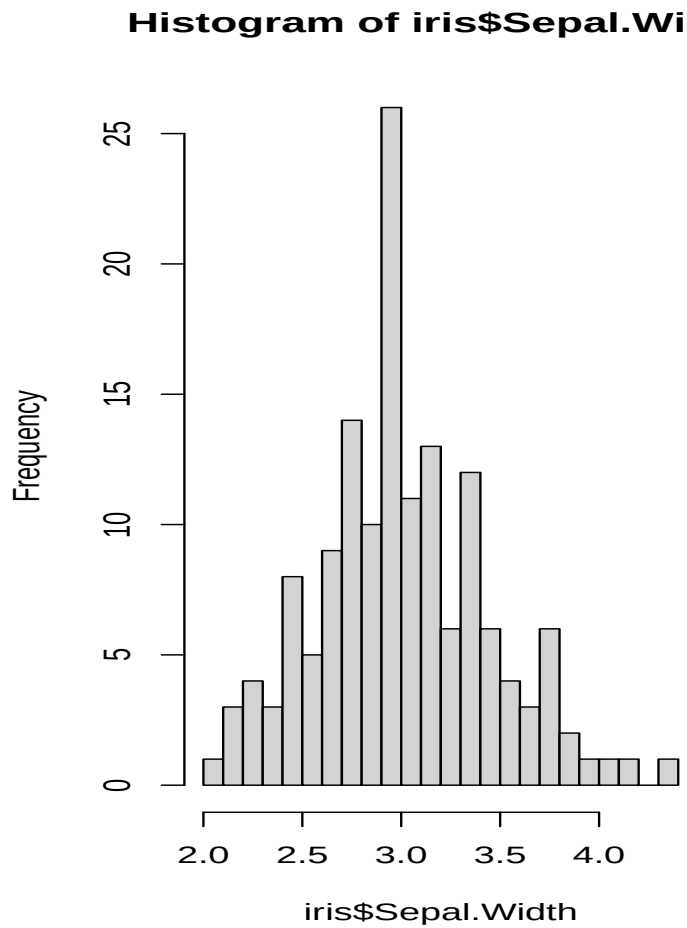
히스토그램 (histogram)

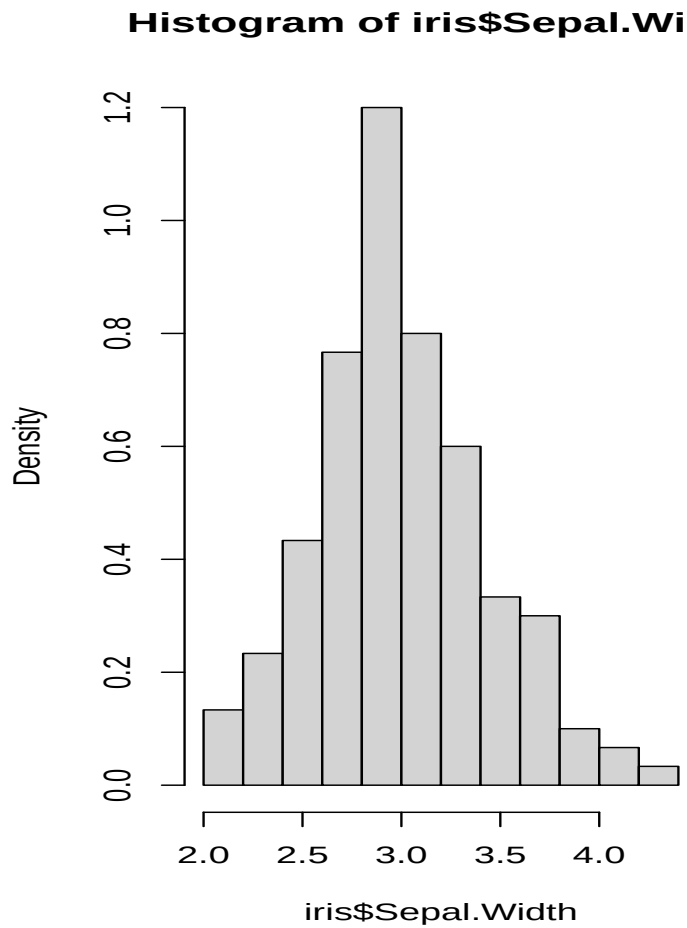
- 계급 구간에 따라 그 위에 넓이가 상대도수 또는 도수에 비례하도록 직사각형을 그린 그림
- 일반적으로 전체 직사각형들의 넓이의 합을 1로 한다.
- 계급을 나누는 방식, 폭의 너비에 따라 모양이 바뀔수 있다.

```
hist(iris$Sepal.Width)
```

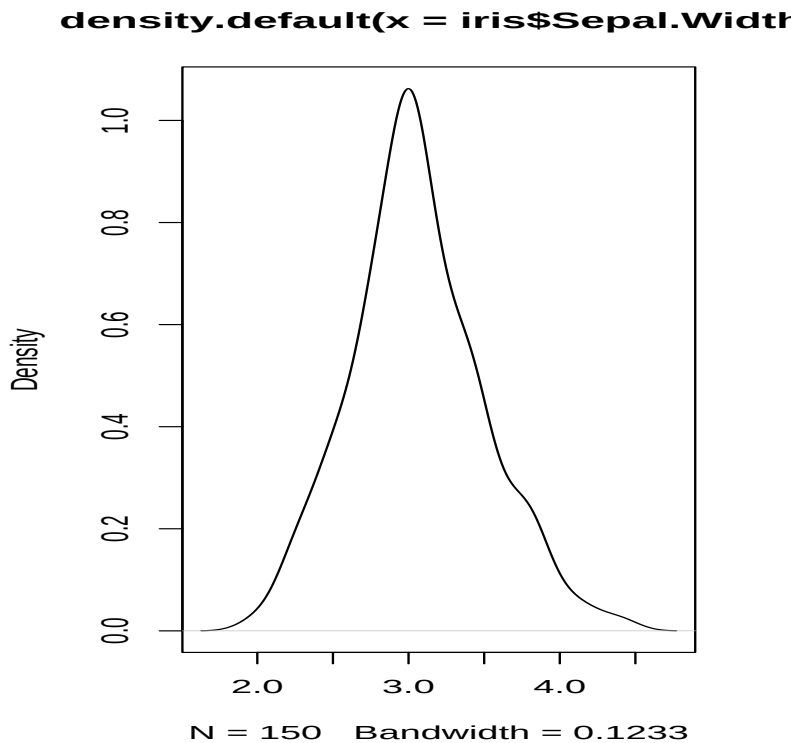


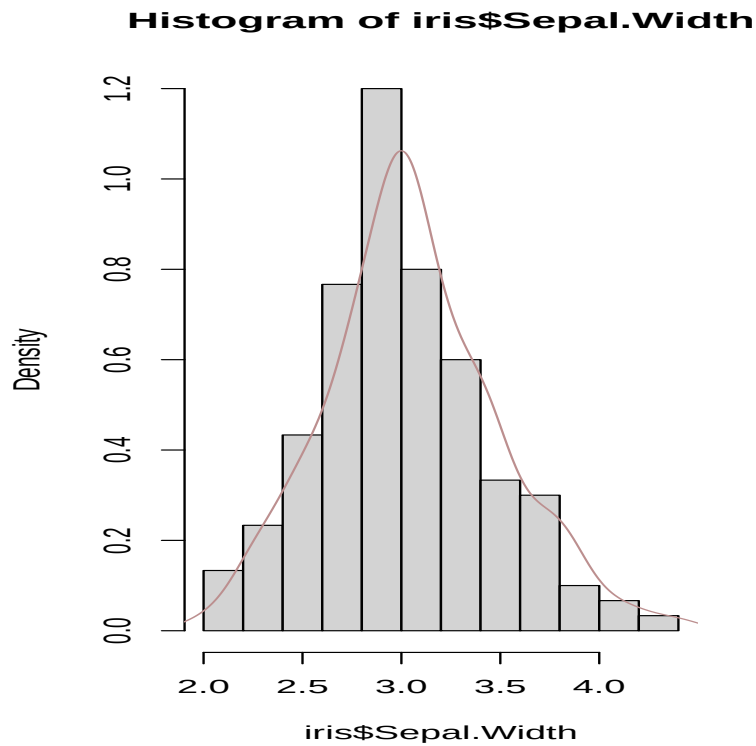
```r
par(mfrow=c(1,2))
hist(iris$Sepal.Width,breaks=20)
hist(iris$Sepal.Width,freq=FALSE)
```





```r
par(mfrow=c(1,2))
plot(density(iris$Sepal.Width))
hist(iris$Sepal.Width,freq=FALSE)
lines(density(iris$Sepal.Width),col='rosybrown')
```





상자 그림 (box plot) 은 몇가지 통계량을 배운 후에 소개.

이차원 자료의 요약

이차원 분할표 (contingency table)

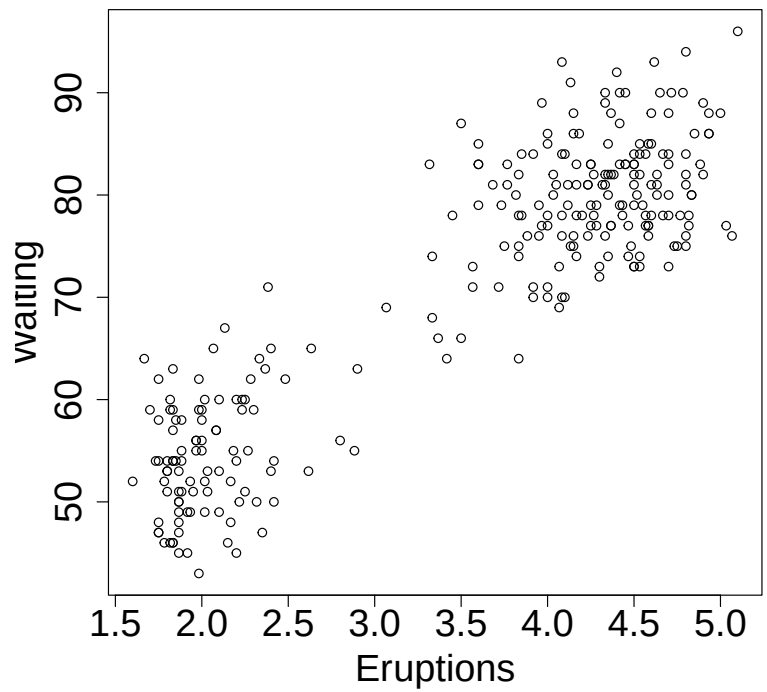
- 각 차원의 특성값의 도수 (또는 상대도수)를 표에 나열한것

Income($)	Job Satisfaction: Dissatisfied	Moderately Satisfied	Very Satisfied	row total
$< 15{,}000$	44	80	82	206
15,000 - 25,000	52	104	120	276
25,000 - 35,000	41	84	115	240
$> 35{,}000$	23	57	91	171
column total	160	325	408	

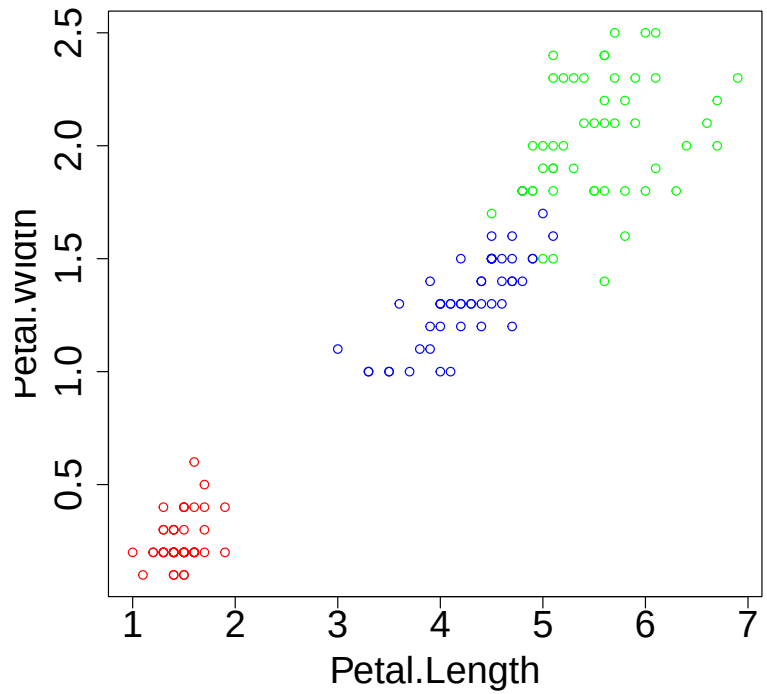
산점도 (scatter plot)

- 두 변수의 변화 관계를 쉽게 알아볼 수 있다.

```
plot(faithful$eruptions,faithful$waiting,xlab='Eruptions',
     ylab='Waiting',cex.lab=2,cex.axis=2)
```



```r
plot(iris$Petal.Length,iris$Petal.Width,xlab='Petal.Length',
     ylab='Petal.Width',cex.lab=2,cex.axis=2,type='n',cex=2)
points(iris$Petal.Length[iris$Species=='setosa'],
       iris$Petal.Width[iris$Species=='setosa'],col='red')
points(iris$Petal.Length[iris$Species=='versicolor'],
       iris$Petal.Width[iris$Species=='versicolor'],col='blue')
points(iris$Petal.Length[iris$Species=='virginica'],
       iris$Petal.Width[iris$Species=='virginica'],col='green')
```



수치로 자료 요약

- 모집단(population) : 자료를 추출할 관심의 대상으로 모든 가능한 추출값들의 집합으로 볼 수 있다.
  - 예) 유권자의 특정 후보 지지도에 대해서 알고 싶다면, 모집단은 전체 유권자들의 지지도 유무의 집합으로 생각할 수 있음.
- 표본(sample) : 일반적으로 모집단을 모두 조사하기 불가능하기때문에 일부만 추출하고 이를 표본이라고 한다.
- 표본 자료를 통해 모집단의 특성을 이해하기를 원함.

- 모수(Parameter): 모집단의 분포의 특징을 나타내는 값
  - 예) 평균, 분산, 표준편차
- 통계량(statistics): 자료를 가지고 계산한 값으로 종류에 따라 모수를 추정하는데 쓰임.
  - 예) 표본평균, 표본분산, 표본표준편차

자료의 중심을 수치로 요약하기

(1) 표본 평균 (sample mean) :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

(2) 중앙값 (median):

자료를 작은값부터 일렬로 나열 했을때, 가운데에 있는 값.
n 개의 자료 $x_1, x_2, \cdots, x_n$ 이 있다고 하자. 작은 값부터 순서대로 나열한 값 (순서 통계량)을 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ 이라고 하자. 중앙값은 n 이 짝수 이면, $(x_{(n/2)} + x_{(n/2+1)})/2$, n이 홀수 이면, $x_{(n/2+1)}$ 이 된다.

(3) 최빈값 (mode) : 자료중 빈도가 가장 높은 자료값

예시 - 간헐천 자료. 첫 15개의 waiting (분출사이의 대기시간)

79 54 74 62 85 55 88 85 51 85 54 84 78 47 83

작은값부터 순서대로 나열

47 51 54 54 55 62 74 78 79 83 84 85 85 85 88

- 중앙값: 78
- 최빈값: 85
- 평균: 70.93

간헐천 전체 자료를 가지고 해보기

```
length(faithful$waiting)
```

```
[1] 272
```

```
mean(faithful$waiting)
```

```
[1] 70.9
```

```
median(faithful$waiting)
```

```
[1] 76
```

```
freq=tabulate(faithful$waiting)
max(freq)
```

```
[1] 15
```

```
which.max(freq)
```

```
[1] 78
```

자료의 산포를 수치로 요약하기

(1) 표본분산 (sample variance)

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

(2) 표본 표준편차 (sample standard deviation)

$$S = \sqrt{S^2}$$

간헐천 자료를 이용해서 자료의 산포도 계산

```
n=length(faithful$eruptions)
sum((faithful$eruptions - mean(faithful$eruptions))^2)/(n-1)  #표본
분산

[1] 1.303

var(faithful$eruptions)  #R에서 표본분산 명령어

[1] 1.303

sqrt(var(faithful$eruptions))

[1] 1.141

sd(faithful$eruptions)  #R에서 표본표준편차 명령어

[1] 1.141
```

(3) 사분위수 범위 (interquartile range, IQR)
$IQR = Q_3 - Q_1$

- p 백분위수 ($p/100$ 분위수): 자료를 크기 순서로 나열했을때 대략 $n \cdot p/100$ 번째 값.

$$\hat{x}_{1-p/100} = \begin{cases} (x_{(k)} + x_{(k+1)})/2 & n \cdot p/100 = k \\ x_{(k+1)} & k < n \cdot p/100 < k+1 \end{cases}$$

- 사분위수 (Quartiles) : 제 i 사분위수: $Q_i = i \cdot 25$ 백분위수, $i = 1, 2, 3$.
  - $Q_2 =$ 중앙값

(4) 표본의 범위 (range)

$$R = x_{(n)} - x_{(1)}$$

```r
pquant=quantile(faithful$eruptions,probs=c(0.25,0.5,0.75))
pquant[3]-pquant[1]  # IQR
```

```
  75%
2.292
```

```r
IQR(faithful$eruptions) # IQR 명령어
```

```
[1] 2.292
```

```r
max(faithful$eruptions)-min(faithful$eruptions) #표본의 범위
```

```
[1] 3.5
```

```r
R=range(faithful$eruptions) # Range를 벡터로 구해주는 명령어
R[2]-R[1]
```

```
[1] 3.5
```

- *IQR*을 이용한 이상치 (Outlier) 찾기

  일반적으로 $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$ 바깥에 있는 자료값들을 의심되는 이상치 (suspected outlier) 로 본다.

- 5개 통계치 요약 (Five-Number Summary)

  (최소값, 1분위수, 중앙값, 3분위수, 최대값)

```r
iqr.val=IQR(faithful$eruptions)
c(pquant[1]-1.5*iqr.val, pquant[3] +1.5*iqr.val)
```

```
   25%    75%
-1.275  7.892
```

```r
faithful$eruptions[faithful$eruptions > pquant[3] +1.5*iqr.val]
```

```
numeric(0)
```

```r
faithful$eruptions[faithful$eruptions < pquant[3] -1.5*iqr.val]
```

```
numeric(0)
```

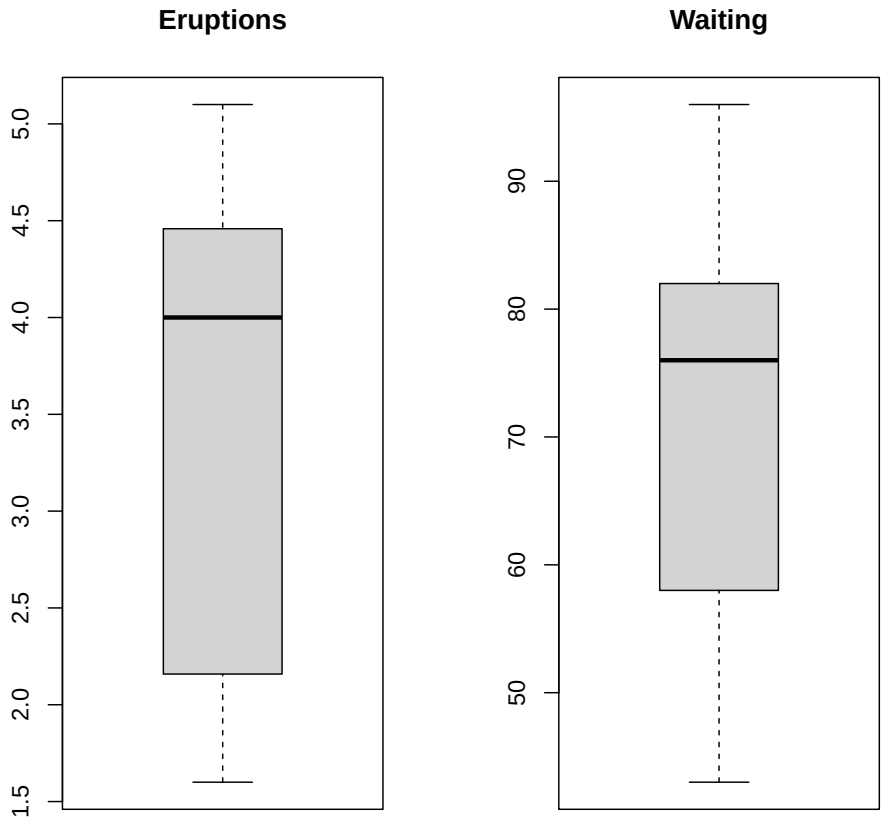
```r
summary(faithful$eruptions)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.60    2.16    4.00    3.49    4.45    5.10
```

상자그림 (box plot)

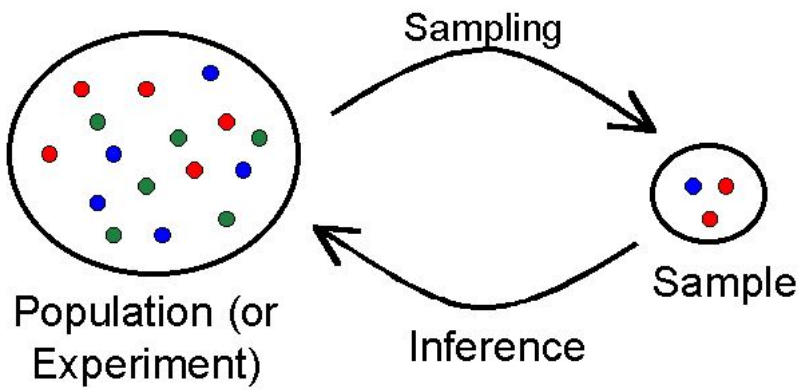
- 5개의 통계치 요약과 의심되는 이상치를 그래프로 표현
- 상자의 아래부터 위까지의 거리는 IQR을 의미, 즉 아래선은 1분위수, 윗선은 3분위수
- 상자의 가운데 선은 중앙값
- 상자로부터 확장된 선의 끝은 R에서는 default로 $1.5IQR$ 를 표시
- 관측값들중에 상자로 부터 $1.5IQR$보다 멀어져 있는 값들을 추가로 표시

```r
par(mfrow=c(1,2))
boxplot(faithful$eruptions,main='Eruptions')
boxplot(faithful$waiting,main='Waiting')
```



통계적 추론

- 표본으로부터의 정보를 이용하여 모집단에 관한 추측이나 결론을 이끌어내는 과정을 통계적 추론(Statistical inference)이라 한다.
- 통계적 추론에서 결론의 신빙성은 표본의 크기에 따라 달라지고, '확률'을 통해 수 값으로 나타난다.
- 통계적 추론은 크게 추정(Estimation)과 유의성 검정(Significance test), 또는 가설 검정(Hypothesis test)으로 나뉜다.



추정

- 모집단의 특성값(모수)에 대한 추측값과 그 오차의 한계를 제시하는 것을 추정이라 한다.
- 모수의 추정은 주로 점추정과 구간추정을 함께 함으로써 이루어진다.

점추정(Point estimation)

모수의 추정에 사용되는 통계량인 **추정량**(Estimator)을
이용하여 표본이 주어지면 그에 따른 값을 제공하는 추정 방식

추정량의 예

- 모평균의 추정량 : 표본평균 $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 모분산의 추정량 : 표본분산 $\hat{\sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

추정량의 평가

추정량을 평가하는 데에도 몇 가지 기준이 있다.

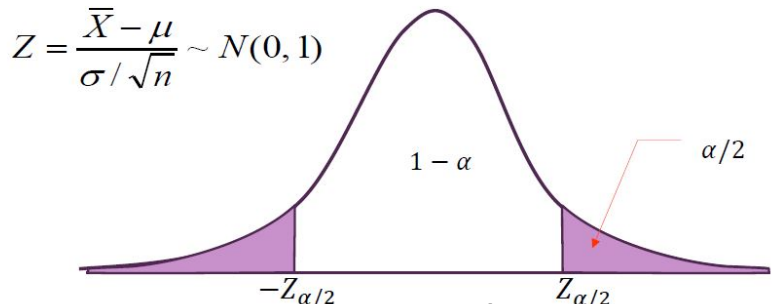
- 모수 θ의 가능한 모든 값에 대해 $E(\hat{\theta}) = \theta$를 만족하는 추정량 $\hat{\theta}$을 불편추정량(Unbiased estimator)이라 한다.
- 예 : 표본평균과 표본분산은 각각 모평균과 모분산의 불편추정량이다.
- 모수 θ의 추정량 $\hat{\theta}$의 표준편차를 추정량의 표준오차(Standard error)라 한다.
- 예 : $X_1, \cdots, X_n$이 모평균이 μ이고 모분산이 σ^2인 모집단에서의 랜덤표본일 때, 표본평균의 표준오차는 다음과 같다.

$$s.e.(\hat{\mu}) = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

구간추정 (Interval estimation)

- 추정량의 신뢰구간 (Confidence interval) 을 제공하여 모수의 범위를 추측하는 것
- 신뢰구간의 형태는 모집단의 분포에 따라 다르다. 정규모집단의 평균을 추정하는 경우, 추정하고자 하는 모평균에 대칭인 형태의 신뢰구간을 사용한다.

예 : 모분산 σ^2 를 알 때 정규모집단의 모평균 μ 의 구간추정



모평균이 가운데 부분에 있을 구간의 확률이 $1\text{-}\alpha$ 이 되는 구간 ($100(1\text{-}\alpha)\%$ 신뢰구간)

$$P(-Z_{\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le Z_{\frac{\alpha}{2}}) = P(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ 를 **오차의 한계**라 하고, $100(1\text{-}\alpha)\%$ 를 **신뢰수준**이라 한다.

신뢰구간의 의미

μ의 $100(1\text{-}\alpha)\%$ 신뢰구간 : 100번의 표본 추출을 통해 얻어진 100개의 신뢰구간

$$(\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

을 관측한 결과, $100(1\text{-}\alpha)\%$ 개 정도의 신뢰구간이 모평균을 포함할 거라 기대되는 신뢰구간

- 예 : n=25, $\sigma = 10$일 때 모평균의 90% 신뢰구간
- $(\bar{X} - Z_{0.05} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{0.05} \frac{\sigma}{\sqrt{n}})$
  $= (\bar{X} - 1.645 \frac{10}{\sqrt{25}}, \bar{X} + 1.645 \frac{10}{\sqrt{25}})$ $(\because Z_{0.05} \simeq 1.645)$
- 표본을 새로 추출할 때마다 $\bar{X}$가 달라지므로 신뢰구간 역시 길이를 유지한 채 정확한 값들은 달라진다.

예시

중앙아메리카의 저소득층 원주민을 대상으로 49 명의 표본조사를 한 결과 혈청 내의 콜레스테롤 양이 평균 $157.02(\mathrm{mg/L})$ 이었다고 한다. 이들 원주민 전체에서 혈청 내의 콜레스테롤 양이 정규분포이고, 표준편차가 $30(\mathrm{mg/L})$라고 할 때, 원주민 전체에서 혈청 내의 콜레스테롤 양의 평균에 대하여 95% 신뢰구간을 구해보자. (단, $Z_{0.025} = 1.96$)

- $n = 49,\ \sigma = 30,\ \bar{x} = 157.02,\ \alpha = 0.05$
- 95% 신뢰구간 : $(\bar{X} - Z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{0.025}\frac{\sigma}{\sqrt{n}}) =$
  $(157.02 - 1.96\frac{30}{\sqrt{49}}\ ,\ 157.02 + 1.96\frac{30}{\sqrt{49}}) = (148.62,165.42)$

```r
n = 49; sigma=30 ; xbar = 157.02 ; alpha=0.05 ; d=5
qnorm(1-alpha/2) # Z_(0.025)

## [1] 1.96

c.i <- c(xbar - qnorm(1-alpha/2)*sigma/sqrt(n),
         xbar + qnorm(1-alpha/2)*sigma/sqrt(n))
c.i

## [1] 148.6 165.4
```

유의성 검정

- 기존의 이론이나 법칙을 부정하는 것으로 보이는 현상이 관측되었을 때, 이 현상의 반증으로서의 강도를 검증함으로써 기존의 이론이나 법칙을 부정하거나 개선할지 결정하게 된다.
- 이 때, 반증을 찾기 위해 설정된 가설(주로 '기존의 가설')을 귀무가설(Null hypothesis, H_0)이라 하고, 귀무가설의 대안으로 상정되는 가설을 대립가설(Alternative hypothesis, H_1)이라 한다.
- 그리고, 귀무가설에 대한 반증의 강도를 제공하는 과정을 유의성 검정(Test of significance)이라 한다.

예시

건물의 소화용으로 사용되는 살수장치가 섭씨 55도에서 작동되도록 제조하려고 한다. 제조공정의 이상 여부를 판단하기 위해 생산품 중에서 표본을 추출하여 작동 시작 온도를 조사하고자 한다. 이러한 조사에서 공정에 이상이 있다는 증거가 뚜렷하면 후속되는 기술적 조치를 하려 한다.

- 살수장치의 평균 작동 시작 온도 : μ
- 공정에 이상이 있는 경우 : $\mu \neq 55$
- $H_0 \ : \ \mu = 55, H_1 \ : \ \mu \neq 55$

만약, 9개의 표본을 관측한 결과 표본평균 $\bar{x} = 55.63$ 이라는 결과가 나왔다면, 어떻게 해석해야 할까?

필요한 용어의 정리

- 모수에 대한 귀무가설 $H_0 : \theta = \theta_0$에 대해 생각해보자.
- 이 때, 대립가설이 $\theta > \theta_0$, 또는 $\theta < \theta_0$와 같이 비교하는 값의 한 쪽에 대해서만 제시되는 가설을 단측(One-sided) 가설이라 하고, $\theta \neq \theta_0$와 같이 양 쪽에 대해서 제시되는 가설을 양측(Two-sided)가설이라 한다.
- 모수에 대한 유의성 검정을 할 때에는, 표본으로부터 얻어진 통계량을 사용하게 된다. 이 때, 검정에 사용되는 통계량을 검정 통계량(Test statistic)이라 한다.

제1종 오류와 제2종 오류

실제현상 / 검정결과	H_0 참	H_1참
H_0채택	옳은 결정	**제 2종 오류**
H_1채택	**제 1종 오류**	옳은 결정

- 귀무가설이 옳은 상황에서 귀무가설을 기각함으로 인해 생기는 오류를 제 1종 오류, 귀무가설이 틀린 상황에서 귀무가설을 기각하지 못함으로 인해 생기는 오류를 제 2종 오류라 한다.
- 유의성 검정에서는 반증의 강도를 '제 1종 오류가 일어날 확률'을 통해 제시한다. 반증의 강도가 높아질 수록 제 1종 오류가 일어날 확률은 낮아지고, 이는 실제 관측된 결과보다 강력한 반증을 얻게 될 확률 역시 낮아짐을 의미한다.

- 이러한 확률을 유의확률(Significance probability), 혹은 P 값(P-value)이라 하고, 유의확률과 비교하여 대립가설의 유의성을 검정하게 될 기준값을 유의수준(Significance level)이라 한다.
- 예컨대, 유의수준이 $\alpha = 0.05$라 함은, 관측 결과보다 더욱 귀무가설에 대한 반증이 강하게 나타날 수 있는 기회가 5% 이하이기를 요구하는 것이다.
- 조사 결과, 유의확률이 지정된 유의수준 이하로 나타나면 **조사 결과가 통계적으로 유의하다**라고 표현할 수 있다.

- 귀무가설 H_0을 기각시킬 수 있는 검정통계량의 관측값의 영역을 기각역(Critical region)이라 한다.
- 유의수준 α 하에서, 유의확률이 유의수준 이하로 나타나게 되는 영역이라고 볼 수 있다.

예 : 정규모집단에서 모평균의 가설검정

H_0 : $\mu = \mu_0$, 검정통계량 $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ under H_0, 관측값 : z_0

H_1	유의확률	기각역
$\mu > \mu_0$	$P = P(Z > z_0)$	$Z > z_\alpha$
$\mu < \mu_0$	$P = P(Z < z_0)$	$Z < z_\alpha$
$\mu \neq \mu_0$	$P = P(\|Z\| > z_0)$	$\|Z\| > z_{\alpha/2}$

대립가설에 종류에 따른 유의확률과 기각역의 형태)

유의성 검정의 절차

- 귀무가설, 대립가설, 유의수준을 설정한다.
- 표본을 추출하고 검정통계량의 값을 계산한다.
- 검정통계량의 값을 유의수준과 비교하여 평가한다.
  이 때, 유의확률이나 기각역을 통해 검정통계량의 값을 평가하고 귀무가설을 기각할 수 있는지 판단하게 된다.(결과는 같다.)
- 가설을 기각할 수 있는지 없는지를 판단하고, 결론을 이끌어낸다.

유의성 검정의 진행 예시

A사에서 생산중인 고양이 사료 캔의 열량은 평균이 $1{,}200\text{kcal}$, 표준편차가 100kcal로 알려져 있다. 이제, 사료의 열량을 늘리기 위해 재료를 일부 변경하여 만든 시제품을 25개 생산하여 조사한 결과 평균 열량이 $\bar{x} = 1240\text{kcal}$이었다. 새로운 재료로 만든 사료 열량의 표준편차가 100kcal로 유지된다고 할 때, 이 조사 결과는 사료의 열량을 늘리기 위한 재료 변경이 성공적임을 뜻하는가? 유의 수준 0.05에서 검정해보자.

유의성 검정의 진행 예시

- 가설 설정 :
  $H_0 \ : \ \mu = 1200, H_1 \ : \ \mu > 1200, \ \sigma = 100, \ n = 25, \ \alpha = 0.05$
- 검정통계량 $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \boldsymbol{N}(0,1)$ under H_0, 관측값
  $z_0 = \frac{1240 - 1200}{100/\sqrt{25}} = 2.0$
- 기각역 : $P(Z \geq z \mid H_0) = 0.05$. 즉, $(Z \geq Z_{0.05} = 1.645)$
  유의확률 : $P(Z \geq z_0 \mid H_0) = 0.0228$
- 기각역을 통한 비교 : $z_0 = 2.0 > 1.645$
  유의확률을 통한 비교 : $0.0228 < 0.05 = \alpha$
- 검정 결과, 귀무가설을 기각할 수 있다.

```
mu0=1200; sigma=100; n=25; alpha=0.05
xbar=1240; z=(xbar-mu0)/(sigma/sqrt(n))
z

[1] 2

z.alpha=qnorm(0.95)
z.alpha

[1] 1.645

pval=1-pnorm(z)
pval

[1] 0.02275
```

모평균에 관한 추론

- 모평균 μ에 관한 추론을 할 때는 일반적으로 표본평균 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 을 사용하게 된다.
- 모분산이 알려진 경우, 정규분포를 통한 통계적 추론을 할 수 있음을 다뤘다.
- 모분산이 알려져있지 않은 경우에는, 표본표준편차를 활용하여 t분포를 통한 추론을 해야 한다.

모평균에 관한 추론(모분산을 모르는 경우)

- 모분산을 아는 정규모집단의 모평균을 추정할 때, 다음과 같은 표준화된 표본평균을 통계량으로 사용한다.
  $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \boldsymbol{N}(0,1)$
- 모분산이 알려져있지 않은 경우에는, 위의 통계량에서 모표준편차 σ 를 표본표준편차 S로 대체한다.
- 이와 같은 과정을 스튜던트화(Studentize)라 하며, 스튜던트화된 표본평균은 다음과 같다.

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

- 이를 활용하여 모평균에 관한 추정이나 유의성검증을 할 수 있다.

모평균 구간추정(모분산을 모르는 경우)

스튜던트화된 표본평균을 활용하여 모분산을 모르는 정규모집단의 평균을 구간추정하는 과정은 다음과 같다. (단, 표본 크기가 충분히 클 경우 정규모집단의 과정은 완화될 수 있고, t분포의 백분위수 대신 표준정규분포의 백분위수가 사용될 수도 있다.)

- $P(-t_{\alpha/2}(n-1) \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)) \; = \; 1-\alpha$
- 위의 식을 모평균 μ에 대해 정리하면 다음과 같다.

$$P(\bar{X} - t_{\alpha/2}(n-1)S/\sqrt{n} \leq \mu \leq \bar{X} + t_{\alpha/2}(n-1)S/\sqrt{n}) = 1-\alpha$$

- 이를 통해 모평균의 $100(1-\alpha)\%$ 신뢰구간을 얻을 수 있다.

$$\left(\bar{X} - t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}\right)$$

예시 : 전구를 생산하는 한 회사에서 현재 생산하는 전구의 평균수명은 1950 시간으로 알려져 있다. 개발 중인 전구의 평균수명 μ가 기존의 전구보다 수명이 더 길다고 할 수 있는지 판단하기 위해 9개의 시제품을 생산하여 그 수명시간을 조사한 결과가 다음과 같다.
$\{\ 2000,\ 1975,\ 1900,\ 2000,\ 1950,\ 1850,\ 1950,\ 2100,\ 1975\ \}$
적절한 가설을 세우고, 수명의 분포가 정규분포라는 전제하에서 가설에 대한 유의수준 5% 의 검정을 해보자. 그리고 유의확률을 구해보자.

- 가설 설정 : $H_0\ :\ \mu = 1950, H_1\ :\ \mu > 1950$
- 유의 수준 : $\alpha = 0.05$, 표본의 수 n=9
- 검정통계량 : $T\ =\ \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t(8)$ under H_0

R Code

```
bulb <- c(2000, 1975, 1900, 2000, 1950, 1850, 1950, 2100, 1975)
mean(bulb)

[1] 1967

sd(bulb)

[1] 69.6

qt(0.95,8) #기각역의 경계값

[1] 1.86
```

```r
t.test(bulb, mu=1950, alternative="greater")
```

```
^^IOne Sample t-test

data:  bulb
t = 0.72, df = 8, p-value = 0.2
alternative hypothesis: true mean is greater than 1950
95 percent confidence interval:
 1924  Inf
sample estimates:
mean of x
     1967
```

대응비교에 의한 모평균의 비교

두 모집단의 평균을 비교할 때 실험단위를 동질적인 쌍으로 묶은 다음, 각 쌍에 두 처리를 임의로 적용하고, 각 쌍에서 모은 관측값의 차로 (처리효과의 차)에 관한 추론을 하는 방법을 대응비교 또는 쌍체비교(Paired comparison)라고 한다.

대응비교의 사례

- 두 종류의 운동화의 밑창의 마모정도 비교 : 피실험자의 왼발, 오른발에 각기 다른 운동화 착용
- 약의 효능검사 : 환자의 초기 증상의 정도에 따라 차이가 나는 것을 통제하기 위해 복용전과 복용후의 효과 측정
- 좌, 우 시력비교

대응비교에 의한 모평균의 비교

- 이질적인 두 실험단위 $X_1, X_2, \cdots, X_n$과 $Y_1, Y_2, \cdots, Y_n$을 비교할 때 동질적인 실험단위로 다음과 같은 쌍을 이루어 자료 구조를 형성한다.(두 실험단위의 표본 수는 당연히 같아야 한다.)

$$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$$

- 이 때, 각각의 처리의 효과를 나타내는 모평균 $\mu_1 = E(X_i), \mu_2 = E(Y_i)$ $(i = 1, 2, \cdots, n)$을 비교하려면, 그 차인 $D_i = X_i - Y_i$를 이용할 수 있다.
- D_i의 모평균은 $E(D_i) = \mu_1 - \mu_2$이므로, 이들의 평균과 표준편차인 $\bar{D}, S_D$를 이용하여 구간추정과 유의성검정을 할 수있다.
- 이러한 추론은 D_i가 정규분포로부터의 랜덤표본이라는 전제 하에 정확한 것이다. (n이 충분히 크면 근사적으로 성립)

모평균의 차이에 관한 구간추정

$\mu_1 - \mu_2 = \delta$의 $100(1-\alpha)\%$ 신뢰구간

- $P(-t_{\alpha/2}(n-1) \leq \frac{\bar{D}-\delta}{S_D/\sqrt{n}} \leq t_{\alpha/2}(n-1)) \;=\; 1-\alpha$
- 위의 식을 모평균의 차이 δ에 대해 정리하면 다음과 같다.

$$P(\bar{D} - t_{\alpha/2}(n-1)S_D/\sqrt{n} \leq \delta \leq \bar{D} + t_{\alpha/2}(n-1)S_D/\sqrt{n}) = 1-\alpha$$

- 이를 통해 모평균의 차이에 대한 $100(1-\alpha)\%$ 신뢰구간을 얻을 수 있다.

$$(\bar{D} - t_{\alpha/2}(n-1) \cdot \frac{S_D}{\sqrt{n}}, \bar{D} + t_{\alpha/2}(n-1) \cdot \frac{S_D}{\sqrt{n}})$$

예시 : 두 종류의 진통제에 대한 상대적 효과의 척도로서, 복용 후 숙면할 수 있는 정도를 비교하려고 한다. 이러한 실험에 참여하기로 한 환자 중에서 소수의 환자를 랜덤추출하여 조사하기로 하였으나 이들 환자들의 건강상태에 상당한 차이가 있음을 알고 있다. 따라서 이들 중 6명의 환자를 랜덤추출하고 각 환자에게 두 종류의 진통제를 각각 1회씩 복용하게 하여 숙면시간의 차이를 이용하여 두 진통제의 효과를 비교하기로 하였다. 이 때, 두 진통제에 의한 숙면시간 차이의 95% 신뢰구간을 구해보자.

(단위 : 시간)

환자	1	2	3	4	5	6
진통제 A	4.8	4.0	5.8	4.9	5.3	7.4
진통제 B	4.0	4.2	5.2	4.9	5.6	7.1
차이	0.8	-0.2	0.6	0.0	-0.3	0.3

자료의 구조

```r
A <- c(4.8,4.0,5.8,4.9,5.3,7.4)
B <- c(4.0,4.2,5.2,4.9,5.6,7.1)
t.test(A-B,mu=0)
```

```
^^IOne Sample t-test

data:  A - B
t = 1.1, df = 5, p-value = 0.3
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.2646  0.6646
sample estimates:
mean of x
      0.2
```

독립 이표본에 의한 모평균의 비교

- 서로 다른 두 모집단을 비교할 때 가장 쉽게 생각할 수 있는 방법은, 두 모집단에서 표본을 각각 랜덤추출하여 얻어지는 서로 독립인 두 표본을 이용하는 것이다.
- 특히, 두 가지의 처리를 비교하는데 두 모집단 간에 동질적인 부분이 없는 경우, 대응비교를 적응하는 데에는 어려움이 따르므로, 실험단위를 두 그룹으로 나누고 서로 다른 처리를 적용하여 그 결과를 비교할 수 있다.
- 예를 들면, 두 가지 교육방법의 효과를 비교하는 경우, 동일한 사람이 두 방법에 모두 적용되기는 어려우므로 대응비교보다는 교육대상자들을 처음부터 두 그룹으로 나누어 그룹별로 교육방법을 달리 하여 비교하는 것이 적절하다. 단, 이 때 두 그룹으로 나누는 과정은 랜덤하게 이루어져야 한다.

이표본에 의한 모평균의 비교 : 모분산을 모를 때 (등분산)

$\mu_1 - \mu_2 = \delta$ 에 관한 유의성 검정
귀무가설 H_0 : $\mu_1 - \mu_2 = \delta_0$ 에 대해 유의수준 α의 유의성 검정을 하려 할 때

- 검정통계량 : $T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$ under H_0

  S_p^2: 합동분산 (pooled variance)
- 검정통계량의 관측값 : $t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{s_p\sqrt{n_1^{-1} + n_2^{-1}}}$
- 이 때 대립가설의 종류에 따른 유의확률과 기각역의 형태는 다음과 같다.

H_1	유의확률	기각역
$\mu_1 - \mu_2 > \delta_0$	$P = P(T > t_0)$	$T > t_\alpha(n_1 + n_2 - 2)$
$\mu_1 - \mu_2 < \delta_0$	$P = P(T < t_0)$	$T < -t_\alpha(n_1 + n_2 - 2)$
$\mu_1 - \mu_2 \neq \delta_0$	$P = P(\|T\| > \|t_0\|)$	$\|T\| > t_{\alpha/2}(n_1 + n_2 - 2)$

대립가설에 종류에 따른 유의확률과 기각역의 형태

예시 : 1972년에 출간된 Science에 마리화나의 주성분과 관련된 실험의 결과가 보고되었다. 이 실험은 마리화나 주성분인 $\Delta^9\text{THC}$와 $11-OH-\Delta^9\text{THC}$가 환각 효과에 미치는 영향의 차이를 알아보기 위한 것이었다.
실험방법 : 건강상태가 비슷한 지원자 12명을 6명씩 랜덤추출하여 두 그룹으로 나눈 후 마리화나의 주성분을 정맥주사 후 환각효과가 느껴지기 시작하는 순간까지의 주사량을 체중 1kg당 10^{-6}g 단위로 측정

- 측정 결과는 다음과 같다.

$\Delta^9\text{THC}$: $\{19.54, 14.47, 16.00, 24.83, 26.39, 11.49\}$
$11-OH-\Delta^9\text{THC}$: $\{15.95, 25.89, 20.53, 15.52, 14.18, 16.00\}$

	Δ^9THC	$11-OH-\Delta^9THC$
표본크기	6	6
평균	18.787	18.012
표준편차	5.908	4.418

표본표준편차의 차이가 크지 않다고 판단, 등분산을 가정한다.

```r
x_1 <- c(19.54, 14.47, 16.00, 24.83, 26.39, 11.49)
x_2 <- c(15.95,25.89,20.53,15.52,14.18,16.00)
# Pooled standard deviation
sqrt(((6-1)*var(x_1)+(6-1)*var(x_2))/(6+6-2))
```

```
[1] 5.217
```

```r
t.test(x_1,x_2,var.equal=TRUE,conf.level=0.95) # t-test procedure
```

```
^^ITwo Sample t-test

data:  x_1 and x_2
t = 0.26, df = 10, p-value = 0.8
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.936  7.486
sample estimates:
mean of x mean of y
    18.79     18.01
```

이표본에 의한 모평균의 비교 : 모분산을 모를 때 (이분산)

모분산을 모르고 등분산 가정을 할 수 없는 상태에서는, 스튜던트화된 표본평균이 따르는 t분포의 자유도를 근사값으로 추정해야 한다. (단, 표본의 크기가 충분히 크면 표준정규분포로 근사할 수도 있다.)

- 일반적으로, 두 정규모집단에서 추출한 표본의 크기 n_1, n_2 가 5 이상이면 스튜던트화된 표본평균의 분포를 다음과 같이 추정할 수 있다.
- $$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t(df^\star)$$

$$df^\star = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{1}{n_1-1}(S_1^2/n_1)^2 + \frac{1}{n_2-1}(S_2^2/n_2)^2}$$

예시 : 질산칼륨의 과다 섭취가 성장을 저해하는 증거가 있는지를 알아보기 위하여 16마리의 쥐를 대상으로 실험을 하였다. 이들 중 9마리를 랜덤추출하여 2000ppm의 질산칼륨을 섭취하게 하고, 나머지 7마리는 일상적인 식사를 하게 하였다. 일정기간 후에 이들의 체중 증가율(%)을 조사한 결과 아래와 같았다.

질산칼륨 섭취군 :
$\{12.7, 19.3, 20.5, 10.5, 14.0, 10.8, 16.6, 14.0, 17.2\}$

규정식 섭취군 : $\{18.2, 32.9, 10.0, 14.3, 16.2, 27.6, 15.7\}$

	질산칼륨 섭취군	규정식 섭취군
표본크기	9	7
평균	15.07	19.27
표준편차	3.56	8.05

질산칼륨의 과다 섭취가 성장을 저해하는 증거가 있는가를 유의수준 5% 에서 검정해보자.

```r
x_1 <- c(12.7,19.3,20.5,10.5,14.0,10.8,16.6,14.0,17.2)
x_2 <- c(18.2,32.9,10.0,14.3,16.2,27.6,15.7)
# t-test procedure
t.test(x_1,x_2,"less",var.equal=FALSE,conf.level=0.95)
```

```
^^IWelch Two Sample t-test

data:  x_1 and x_2
t = -1.3, df = 7.8, p-value = 0.1
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  -Inf 1.887
sample estimates:
mean of x mean of y
    15.07     19.27
```

정리

- 자료를 정보로 변환하는데 시각화가 도움이 된다.
- 다양한 시각화 방법이 존재한다 - GIS이용, word cloud 등
- R에서 가장 기본적인 plot() 사용법을 다루었다.
- 기술통계에서의 수치요약 (평균, 중앙값 최빈값, 분산, 표준편차)과 그래프요약 (막대그래프, 원그래프, 히스토그램, 상자그림)을 통해 자료를 요약할 수 있다.
- 통계적 추론은 표본(자료)를 이용하여 모집단의 특징을 나타내는 모수를 추정하고 검정하는것으로, 표본평균을 이용한 모집단의 평균에 대한 추정과 검정을 배웠다.

참고자료

- Norman Matloff, The Art of R programming
- Jeffrey Stanton, Introduction to Data Science
  http://jsresearch.net/wiki/projects/teachdatascience
- 일반 통계학, 서울대학교 통계학과