통계 분석

2019년 2학기

강봉주

통계학이란?

통계학 이란?

통계학이란 단어는 라틴어의 “statisticum collegium” 즉, 국가평의회와 이탈리아어인 “statista” 즉, 정치가로부터 도출되었다고 한다.

18세기 중반에 독일의 갓프라이드 에이큰월(Gottfried Aschenwall)이 처음으로 “statistik" 이라는 단어를 사용하였으며 이를 국가의 과학(science of state) 즉, 국가에 대한 데이터를 분석하는 것이라고 정의하였다.

이후 19세기 초까지는 일반적으로 데이터의 수집과 분류의 의미로 사용되었으며 영어로는 18세기 후반에 존 싱클레어 경이 “Statistical Account of Scotland” 책에서 등장하였다.

통계학 이란?

Wikipedia: Statistics is a branch of mathematics working with data collection, organization, analysis, interpretation and presentation

위키피디아: 통계학은 수학의 한 분야로써 데이터 수집, 정리, 분석, 해석 및 표현을 연구하는 분야이다.

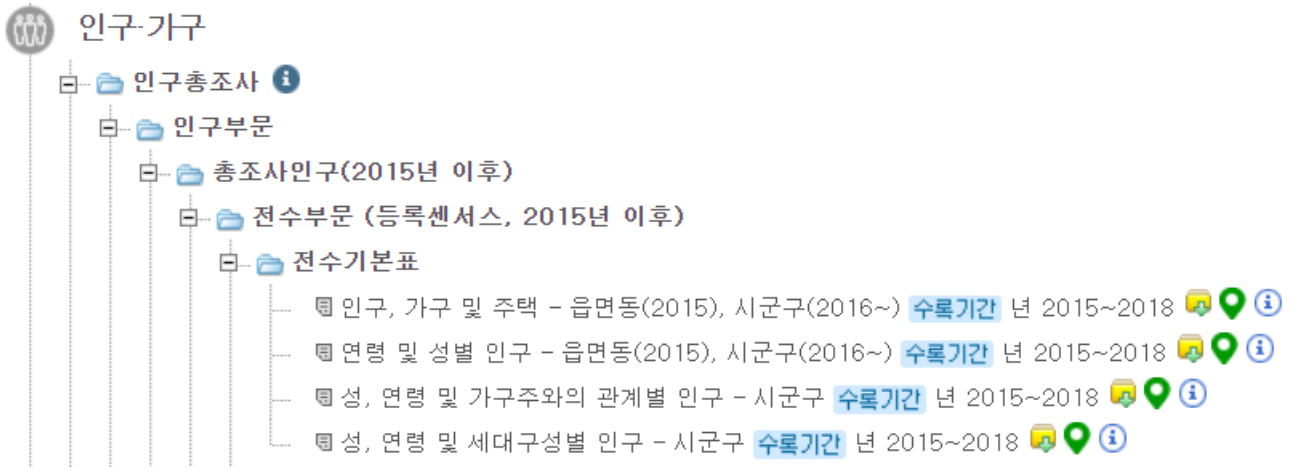
통계학 이란?

통계학의 유명한 학자인 R.A. Fisher의 [SMRW]의 문헌에서는 "통계학은 응용수학의 한 분야이며 **관측 데이터에 적용하는 수학**이다"이라고 정의하고 있다. 더 나아가 통계학은 모집단(population), 변동(variation), 데이터 축소(reduction) 방법에 대한 연구로 정의하고 있다.

통계학의 정의에서 보면 공통적으로 관측된 데이터를 기반으로 하고 있다. 이때 관측된 데이터는 어떤 것을 의미하는 것일까?

통계학 이란?

http://kostat.go.kr/portal/korea/index.action



통계학 이란?

http://kostat.go.kr/portal/korea/index.action

자료갱신일 : 2019-08-30 / 수록기간 : 년 2015 ~ 2018 / 자료문의처 : 042-481-3756

일괄설정 + | 항목[26/26] | 행정구역별(시군구)… | 성별[3/3] | 연령별[17/17] | 시점[1/4] | 통계표조회

(단위 : 명) | 새창보기 | 주석 | URL

행정구역별(시군구)	성별	연령별	2018				
			일반가구원	1세대가구-계	1세대가구-부부	1세대가구-부부+미혼형제자매	1서+기
전국	계	합계	48,664,922	7,305,520	6,494,060	57,695	
		15세미만	6,465,942	6,442	X	16	
		15~19세	2,410,324	19,921	378	105	
		20~24세	2,722,228	91,037	10,801	816	
		25~29세	3,204,366	273,313	145,394	4,198	
		30~34세	3,076,625	424,862	334,534	7,103	
		35~39세	3,912,833	344,671	275,122	6,227	
		40~44세	3,766,795	251,601	194,983	4,752	
		45~49세	4,346,252	399,791	335,209	6,092	
		50~54세	4,052,597	595,162	530,984	6,428	
		55~59세	4,187,673	1,025,849	951,636	7,666	
		60~64세	3,379,781	1,235,386	1,168,211	6,708	
		65~69세	2,318,804	996,778	955,949	3,876	
		70~74세	1,796,606	753,945	728,715	2,240	
		75~79세	1,536,930	554,717	539,844	1,049	
		80~84세	938,037	251,684	245,181	335	
		85세이상	549,129	80,361	77,119	84	

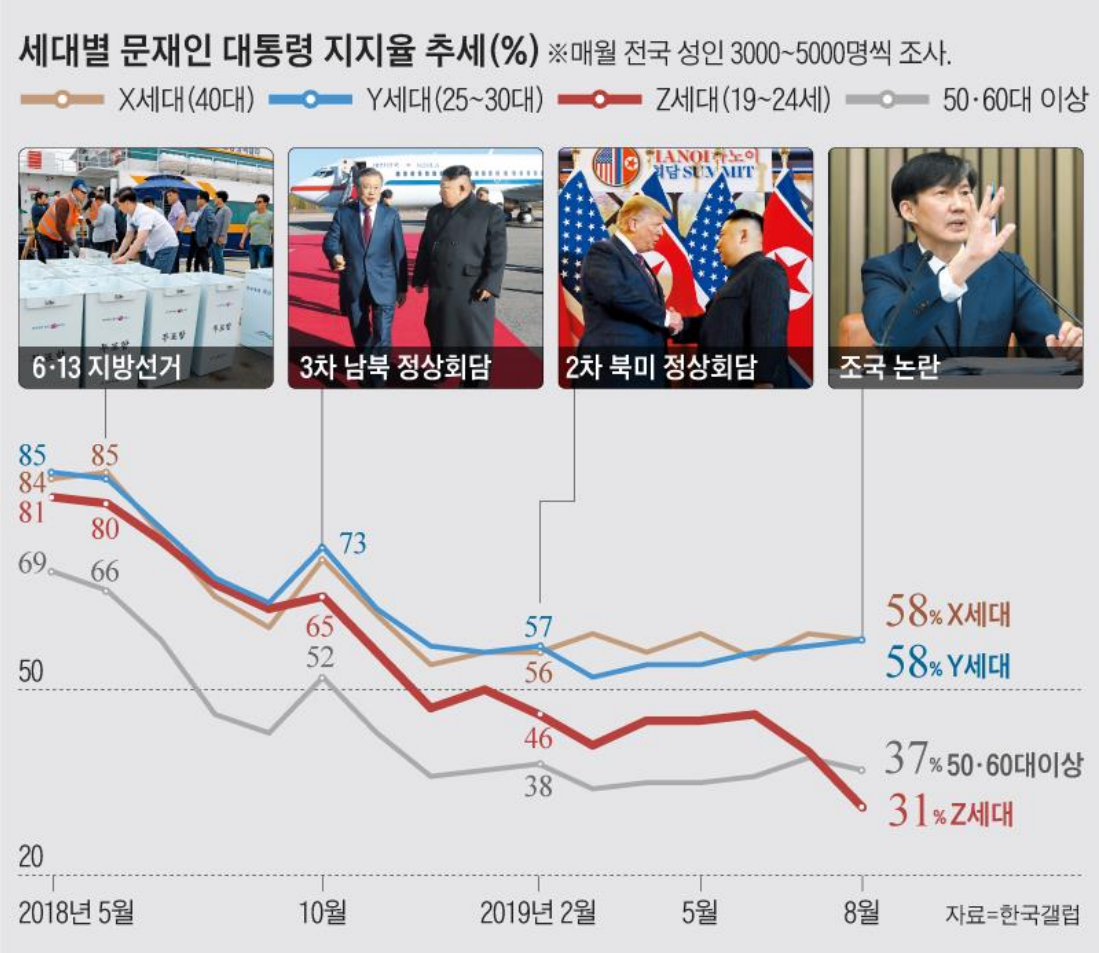
통계학 이란?

https://www.nesdc.go.kr/portal/main.do

등록번호	조사기관명	조사의뢰자	여론조사명칭	등록일	지역	결정사항
6069	(주)한국갤럽조사연구소	한국갤럽 자체 조사	전국 정기(정례)조사 국회의원선거 대통령선거 정당지지도 2019년 9월 1주	2019-09-06	전국	-
6068	케이에스오아이 주식회사 (한국사회여론연구소)	한국사회여론연구소 (KSOI) 자체조사	전국 정기(정례)조사 정당지지도 , 국정운영평가 및 주요현안 등	2019-09-05	전국	-
6067	(주)리얼미터	tbs	전국 정기(정례)조사 정당지지도 9월 1주차 주중집계	2019-09-04	전국	-
6066	(주)리서치뷰	UPI뉴스 &UPI뉴스+	전국 정기(정례)조사 정당지지도 국회의원선거 21대 총선 지역구, 비례대표 정당지지도 등	2019-09-03	전국	-
6065	(주)리얼미터	오마이뉴스	전국 정기(정례)조사 대통령선거 차기대선주자	2019-09-02	전국	-
6064	(주)조원씨앤아이	쿠키뉴스	전국 정기(정례)조사 정당지지도 대통령선거 차기대선지지도	2019-09-02	전국	-
6063	(주)유앤미리서치	대구신문	전국 국회의원선거 정당지지도	2019-09-02	전국	-
6062	(주)디오피니언	내일신문	전국 정기(정례)조사 정당지지도 국회의원선거	2019-09-02	전국	-

통계학 이란?

http://www.gallup.co.kr/



통계학 이란?

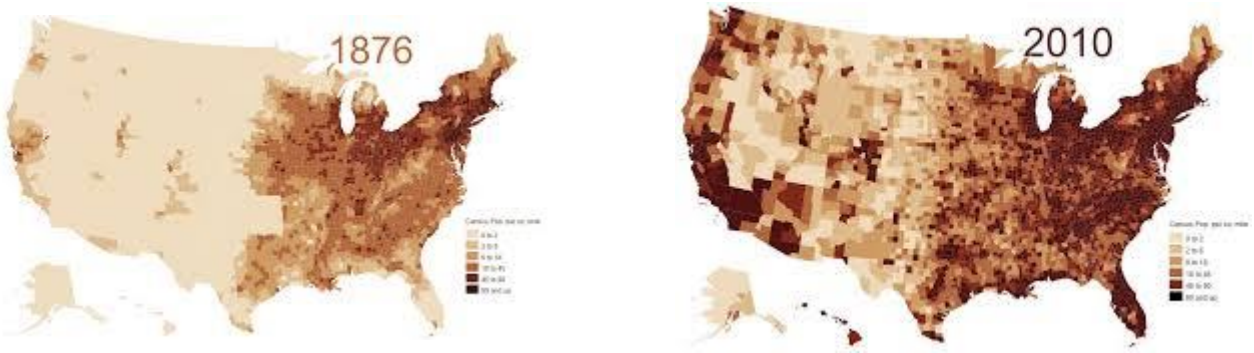
메모리 칩을 생산하기 위한 아주 많은 공정과 거기에서 발생하는 데이터

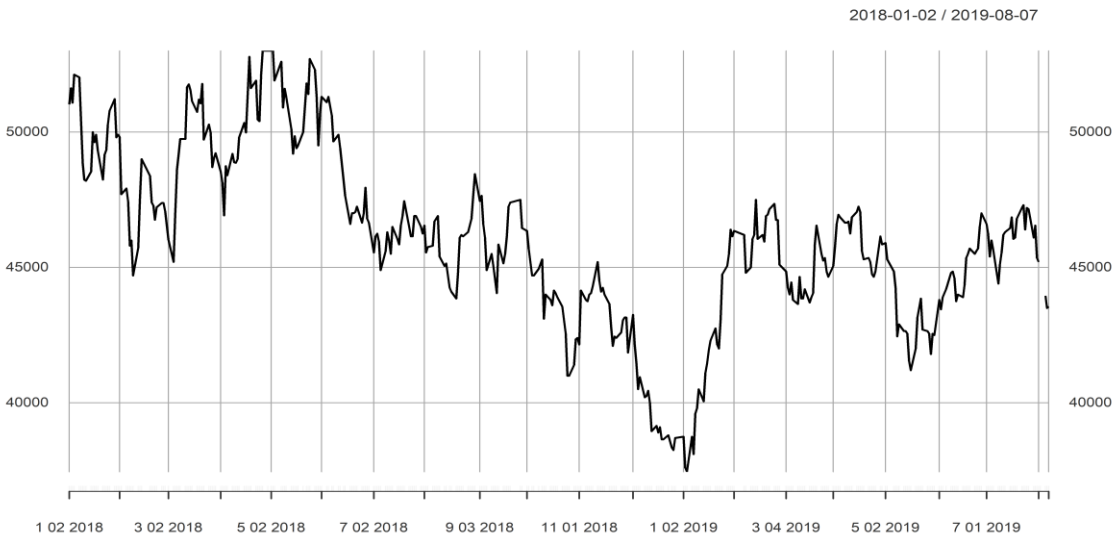
통계학 이란?

데이터를 수집하는 목적은 대상이 되는 한 줄의 데이터인 관측값 들을 수집하였을 때 얻고자 하는 그 무엇이다. 즉, 우리가 모르는 모집단의 특성을 기술(description)하거나 추론(inference)하고자 하는 것이다.

모집단(population)은 기 존재하는 객체들의 집합이거나 경험으로부터 상상할 수 있는 무한 객체 집합 가령, 동전 던지기를 무한히 반복할 때 앞면과 뒷면이 생길 수 있는 조합과 같이 정의할 수 있다. 무한 객체 집합의 예는 가령 삼성전자 주가가 3년 후에 금액의 예상 가능한 결과 집합 등도 하나의 예이다.

통계학 이란?


1876
2010

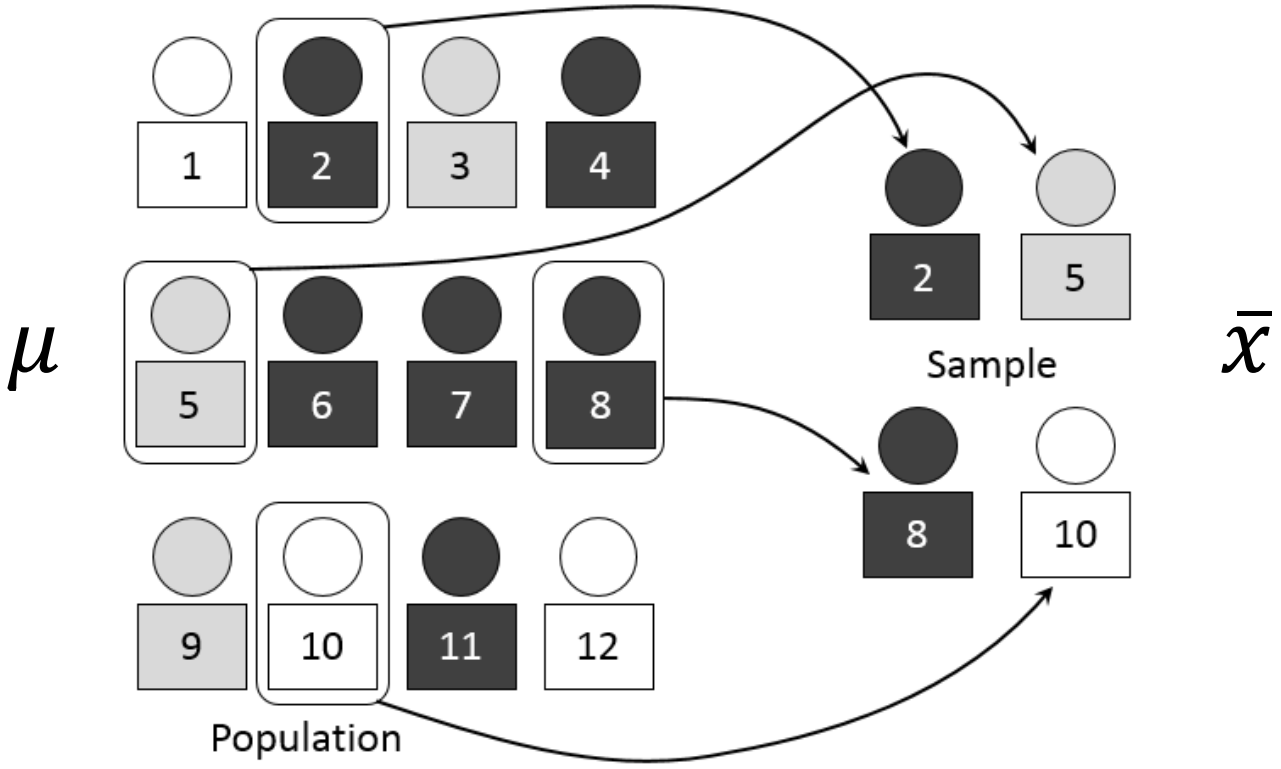

2018-01-02 / 2019-08-07
50000
45000
40000
1 02 2018
3 02 2018
5 02 2018
7 02 2018
9 03 2018
11 01 2018
1 02 2019
3 04 2019
5 02 2019
7 01 2019

통계학 이란?

Yahoo Finance 에서 2008년 1월 1일 부터 현재까지의 종가 기준의 삼성전자 시계열 도표(time series plot)를 그려보자.

통계학 이란?

통계적 표본추출 또는 표집(statistical sampling): 관심 대상인 모집단의 특성을 반영하도록 추출 또는 생성



통계학 이란?

통계적 표본추출 또는 표집(statistical sampling): 관심 대상인 모집단의 특성을 반영하도록 추출 또는 생성

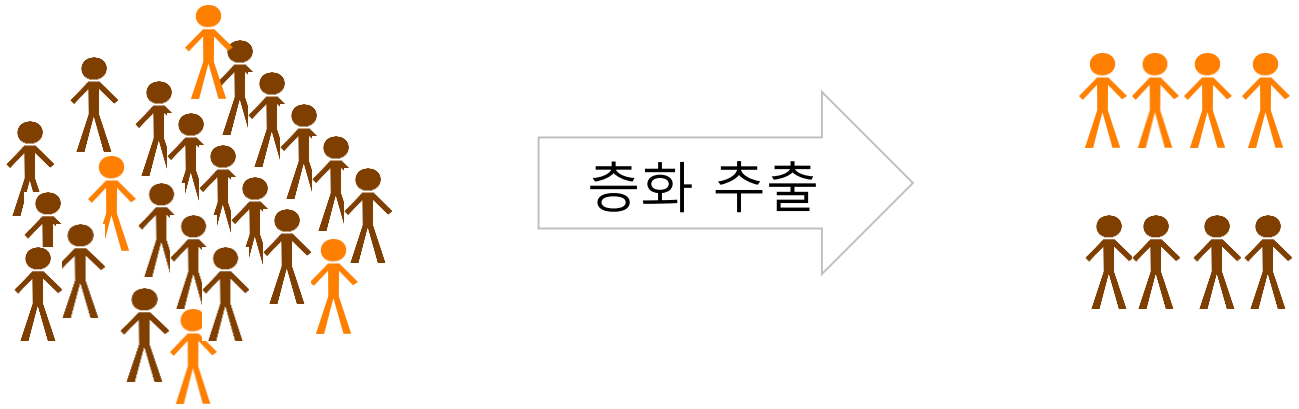
```
# 임의 표본
set.seed(0)
pop <- rnorm(12)
mean(pop)

smp_no = sample(1:12, 4)
mean(pop[smp_no])
```

통계학 이란?

보험 사기 적발 시스템

- 보험 사기의 예가 매우 작음
- 보험 사기의 경우는 전부 추출
- 보험 사기가 아닌 경우는 일부 추출
- 일종의 층화 추출(stratified random sampling)



통계학 이란?

다음의 데이터에서 1 그룹은 전부 추출하고 0 그룹은 1그룹과 크기가 같은 표본을 추출한다.

```
set.seed(123)
pop <- data.frame(
   group = c(rep('1', 10), rep('0', 100)),
   x = runif(110)
)
```

통계학 이란?

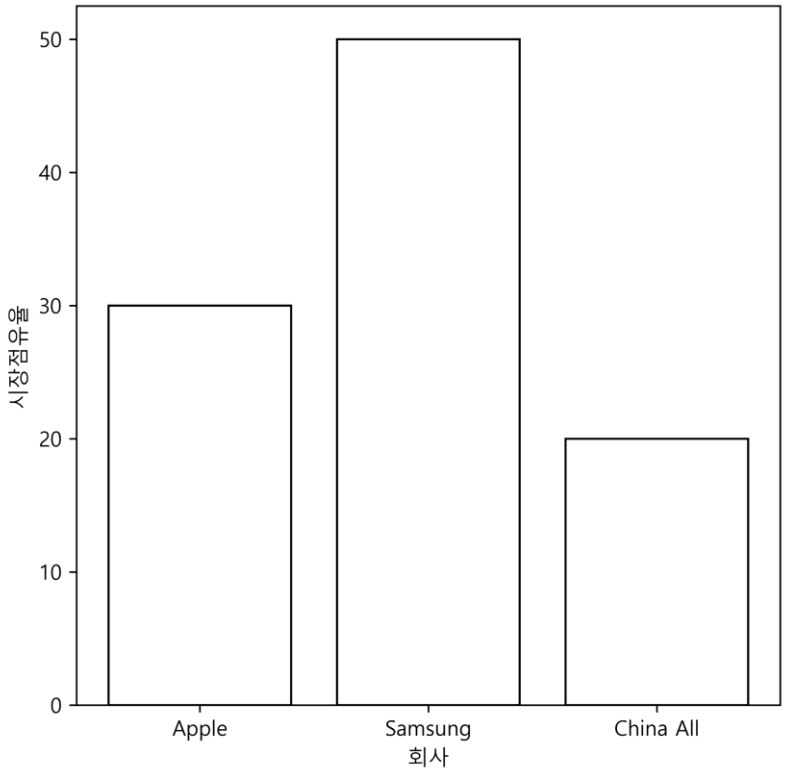
[기술 통계 분석: Descriptive Statistics]

- 하나의 기술 통계량(statistic)은 데이터의 요약 통계량
- 기술 통계 분석은 기술 통계량을 이용하고 분석하는 과정
- 주로 중심적인 경향과 산포를 파악
- 모집단과 표본에 둘 다 적용 가능
- 중심과 산포를 파악할 수 있는 각 종 그래프

$$\bar{x}, \sigma$$

통계학 이란?

[기술 통계 분석: Descriptive Statistics]



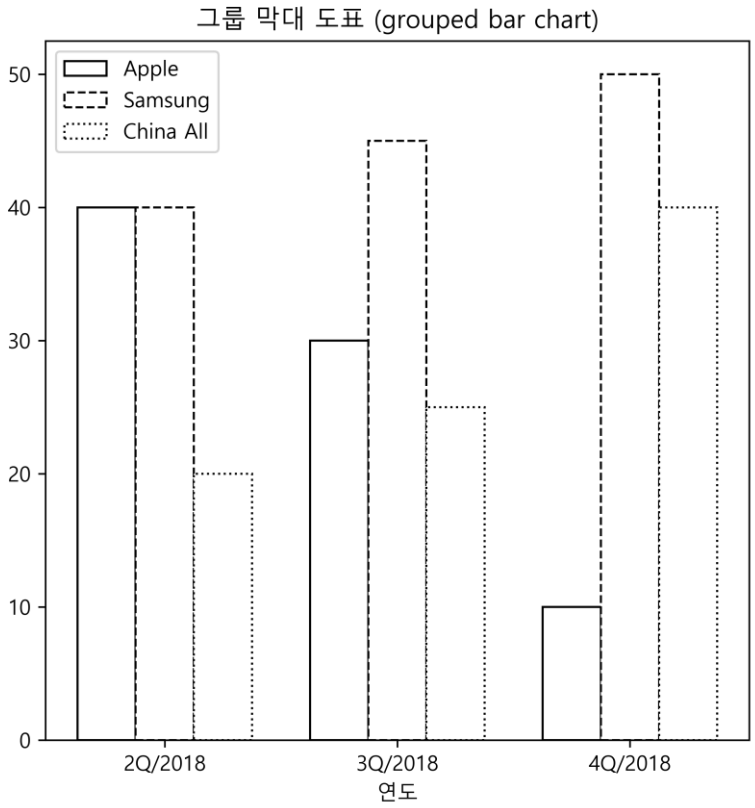
통계학 이란?

[기술 통계 분석: Descriptive Statistics]

```
plot_data <- data.frame(
    names=c('Apple', 'Samsung', 'China
All'),
    bar_height=c(30, 50, 20)
)
```

통계학 이란?

[기술 통계 분석: Descriptive Statistics]



통계학 이란?

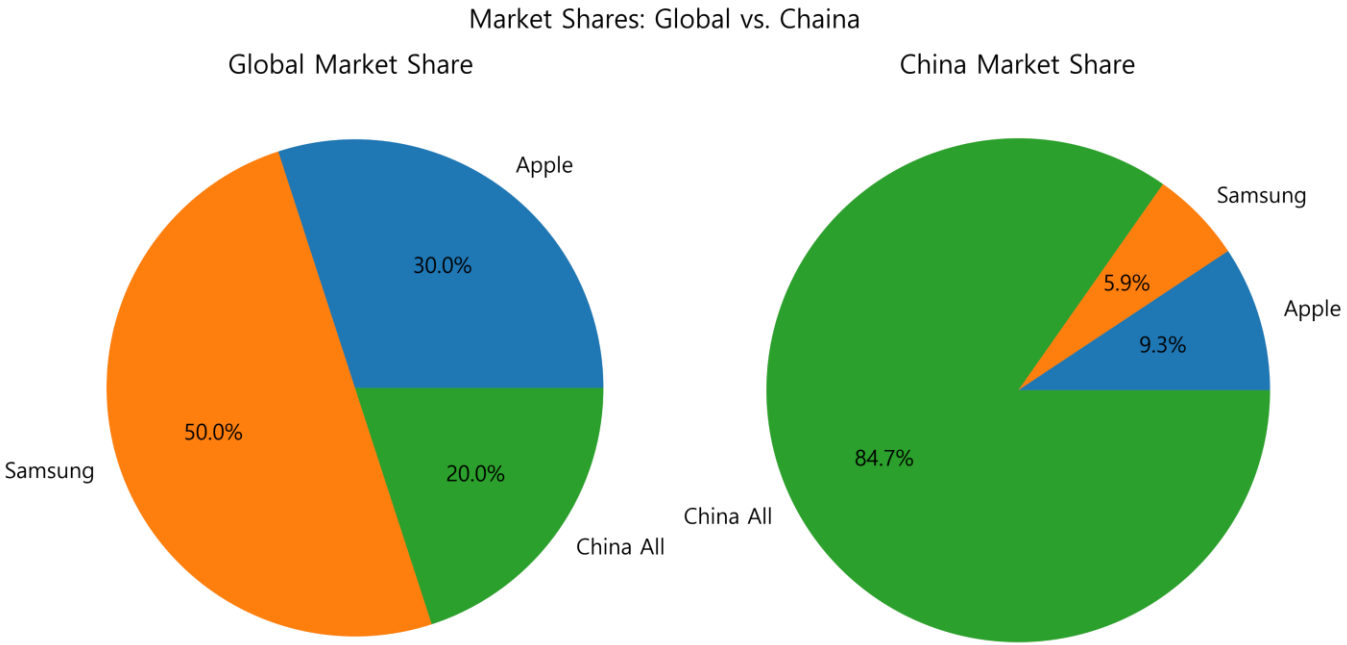
[기술 통계 분석: Descriptive Statistics]

```
plot_df <- data.frame(
    group = rep(c('Apple', 'Samsung','China'), 3),
    quarter = rep(c('2Q/2018', '3Q/2018', '4Q/2018'), each=3),
    count = c(40, 40, 20, 30, 45, 25, 10, 50, 40)
)

# 벡터를 인자로 인코딩: 가만 두면 자동으로 정렬함
plot_df$group <- factor(plot_df$group,
                        levels = c('Apple', 'Samsung','China'))
```

통계학 이란?

[기술 통계 분석: Descriptive Statistics]



통계학 이란?

[기술 통계 분석: Descriptive Statistics]

```
plot_df <- data.frame(
    group = rep(c('Global Market Share', 'China Market Share'),
                each=3),
    labels = rep(c('Apple', 'Samsung', 'China All'), 2),
    sizes = c(30, 50, 20, 11, 7, 100)
    )
```

통계학 이란?

[기술 통계 분석: Descriptive Statistics]

통계학 이란?

[기술 통계 분석: Descriptive Statistics]

```
library(quantmod)

# 삼성전자의 야후 코드
# https://finance.yahoo.com/quote/005930.KS/

symbol <- c("005930.KS")
df <- getSymbols(symbol, src="yahoo",env = NULL,
                 from=as.Date("2014-01-01"))
```

통계학 이란?

[기술 통계 분석: Descriptive Statistics]

통계학 이란?

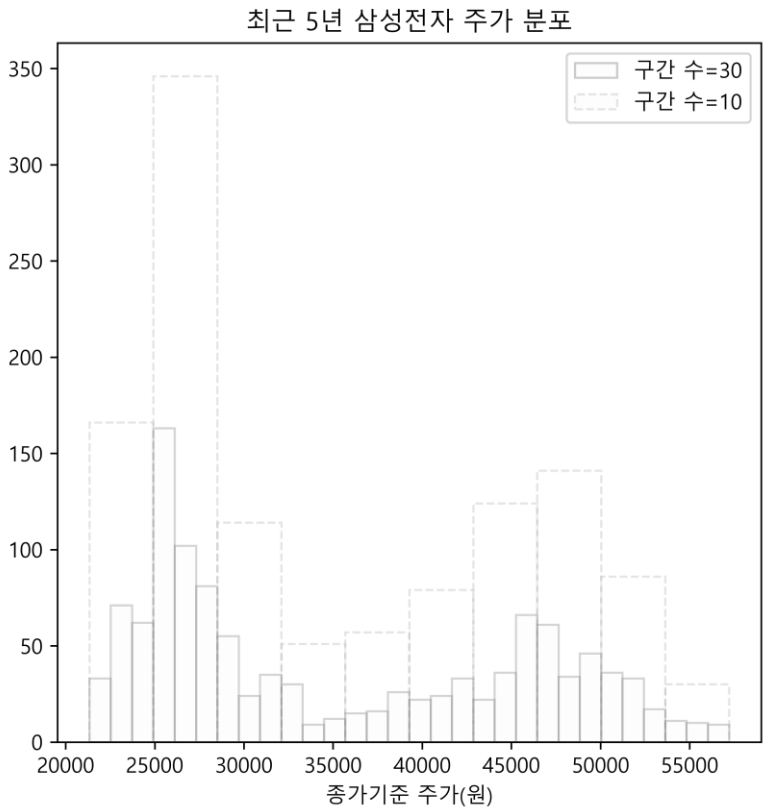
[기술 통계 분석: Descriptive Statistics]

```
# LG화학의 야후 코드
# https://finance.yahoo.com/quote/051910.KS/

symbol <- c("051910.KS")
df2 <- getSymbols(symbol, src="yahoo",env = NULL,
                  from=as.Date("2014-01-01"))
```

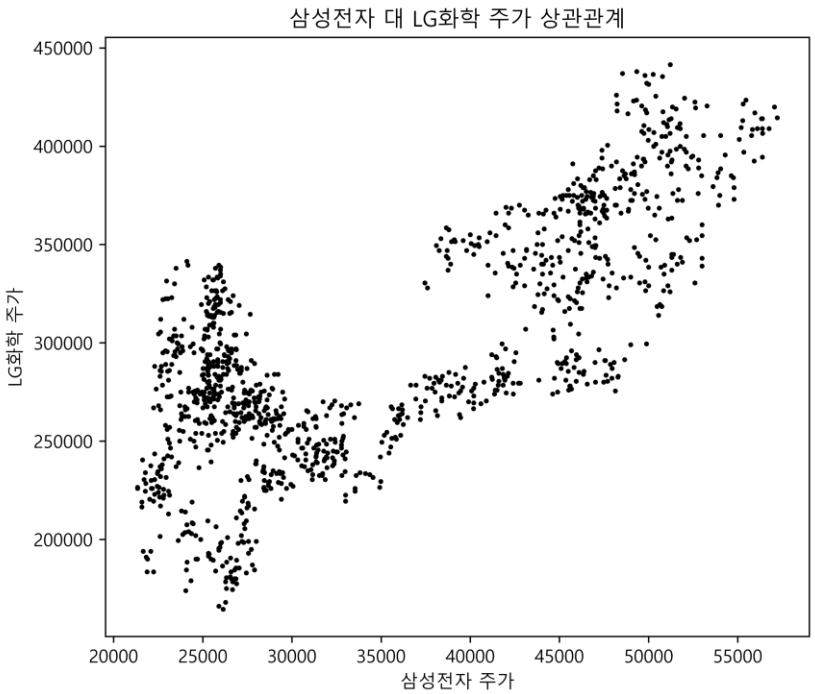
통계학 이란?

[기술 통계 분석: Descriptive Statistics]



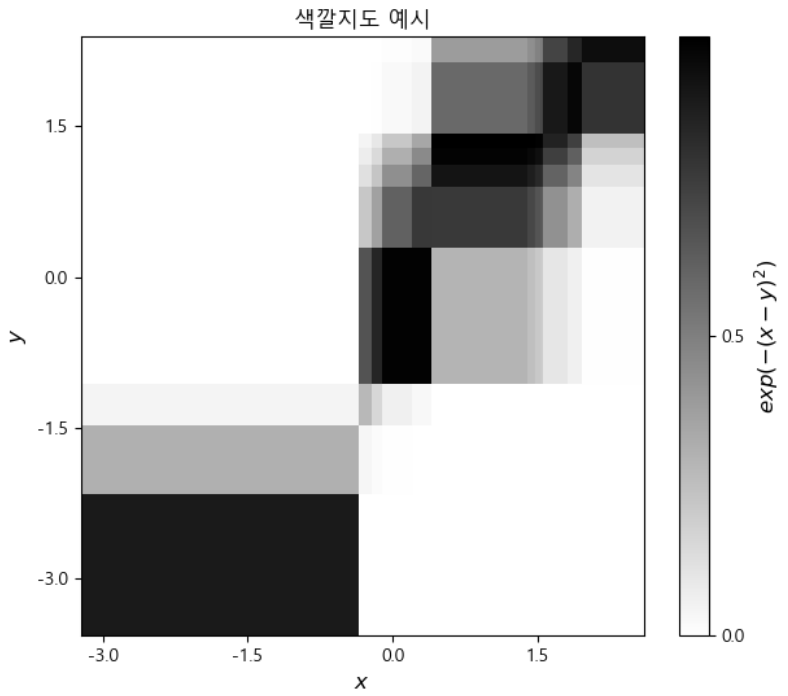
통계학 이란?

[기술 통계 분석: Descriptive Statistics]



통계학 이란?

[기술 통계 분석: Descriptive Statistics]



통계학 이란?

[기술 통계 분석: Descriptive Statistics]

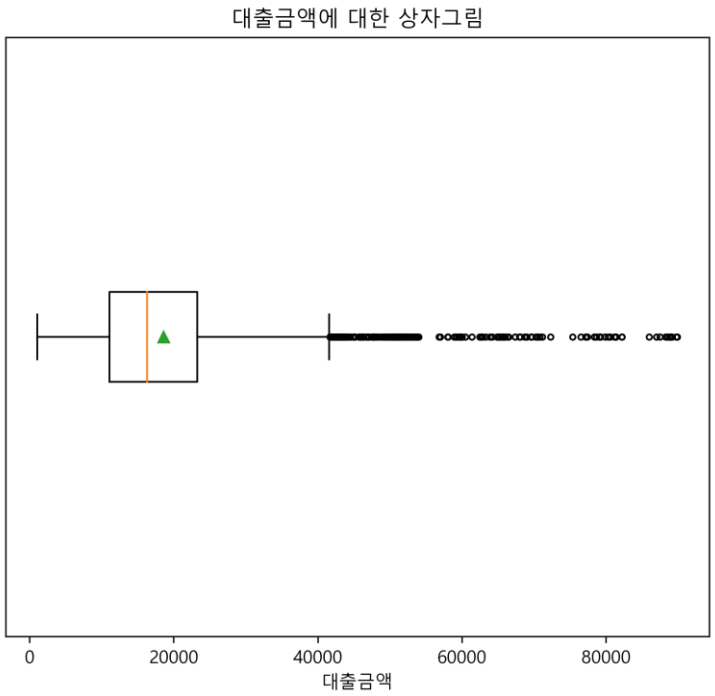
```
# x, y 각 값의 생성
set.seed(123)
x <- rnorm(n=200)
set.seed(1234)
y <- rnorm(n=200)

# x, y 모든 조합에 대한 값 생성
zfunc <- function(x,y){
    return(exp(-(x-y)**2))
}

grid <- expand.grid(X=x, Y=y)
grid$zvalues <- zfunc(grid$X,grid$Y)
```

통계학 이란?

[기술 통계 분석: Descriptive Statistics]



통계학 이란?

[기술 통계 분석: Descriptive Statistics]

```
# 데이터 구성
library(data.table)
df <- fread('https://github.com/bong-ju-kang/kmu-mba-statistics/raw/master/Data/hmeq.csv')
```

통계학 이란?

[기술 통계 분석: Descriptive Statistics]

현재까지 설명된 모든 그래프에 대하여

1) 데이터를 생성한다.
2) 그래프를 그려본다.

통계학 이란?

[통계분석 기법]

데이터에 대한 즉 표본에 대한 이해를 하고 난 후 모집단의 특성 즉, 구체적으로는 분포에 대한 추론(inference)을 하게 되는데 어떠한 통계적인 기법들이 적용될 수 있을까?

통계학 이란?

[통계분석 기법]

데이터에 대한 즉 표본에 대한 이해를 하고 난 후 모집단의 특성 즉, 구체적으로는 분포에 대한 추론(inference)을 하게 되는데 어떠한 통계적인 기법들이 적용될 수 있을까?

통계학 이란?

[데이터 목록]

데이터 이름	데이터 원천
[BANK]	https://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip
[HOUSING]	https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data
[HAS]	통계청 마이크로 데이터: 가계자산조사 > 연간자료(제공)[2006년 가계자산조사]
[HFWS]	통계청 마이크로 데이터: 가계금융복지조사(2017년 이후) > 가구마스터(제공)[2018가구금융복지조사]