기술통계
Descriptive Statistics

CE730 통계와 금융

조남운

Namun Cho/ mailto:namun@snu.ac.kr

주제

- 자료의 구분
- 자료의 정리
- 표본자료의 대표치

자료의 구분

Namun Cho/ mailto:namun@snu.ac.kr

자료의 분류

분류 기준	자료분류
숫자가 의미있는가?	양적 자료 / 질적 자료
셀 수 있는 숫자인가?	이산형 자료 / 연속형 자료
자료의 측정과 시간	횡단면 자료 / 종단면 자료 / 패널 자료

Namun Cho/ mailto:namun@snu.ac.kr

양적 / 질적 자료
Quantitative / Qualitative

- Quantitative Model은 수식, 혹은 논리적 관계로 표현 가능
  - Quantitative Data에서의 수는 의미가 있음
  - 연산 가능
  - 예: 가계소득, 길이, 학점
- Qualitative Model
  - Qualitative Data에서의 수는 식별 이상의 의미가 없음
  - 연산 무의미
  - 명목형 자료, 순서형 자료
  - 예: 성별, 국적, 혈액형

이산형 / 연속형 자료
Discrete / Continuous

- 셀 수 있는 자료
  - 예: 주사위의 눈금, 도산 기업체의 수 등
- 셀 수 없는 자료
  - 예: 질량, 길이, (사실상) 가계소득

횡단면 / 종단면 / 패널 자료 Cross-sectional / Longitudinal / Panel

- 횡단면 자료 Cross-sectional Data
  - 고정된 시간에 획득한 자료
  - 예: 특정 사안에 대한 설문조사
- 종단면 자료 Longitudinal Data or, Time Series Data
  - 시간의 흐름에 따라 측정한 자료
  - 예: 국내총생산, 실업률, 물가지표
- 패널 자료 Panel Data
  - 주기적으로 동일 표본으로부터 횡단면자료를 획득
  - 횡단면과 종단면이 모두 존재
  - 예: 재정패널, 복지패널 등

척도
Scales

- 수집된 자료의 기준설정을 위한 분류
  - 명목척도 Nominal Scales
  - 순서척도 Ordinal Scales
  - 구간척도 Interval Scales
  - 비율척도 Ratio Scales

Namun Cho/ mailto:namun@snu.ac.kr

명목척도
Nominal Scales

- 각 범주의 숫자는 오로지 식별만을 위해 기능함
- 코드 (code)
  - 예:
    - 성별코드 0:여성, 1:남성
    - 지역코드 02:서울, 031:경기도 ...
    - 결혼여부 0:미혼 1:기혼
- 숫자 내용은 의미 없음

순서척도
Ordinal Scales

- 명목척도와 마찬가지로 숫자의 크기는 무의미
- 다만 숫자의 크기 순서는 의미가 있음
- 예: 리커르트 척도 (5점척도)
  - 1:전혀 그렇지 않다 2: 그렇지 않다 3: 중립 4: 그렇다 5: 매우 그렇다
  - 1: 초등학교 졸 2: 중학교졸 ..

구간척도
Interval Scales

- 등급간 서열 뿐만 아니라 간격이 일정
- 더하거나 뺄 수 있음
- 예: 소득 수준
  - 1: 1000만원 미만
  - 2: 1000 - 2000만원
  - …
  - 4: 3000 - 4000만원
  - 5: 4000만원 이상

비율척도
Ratio Scales

- 구간형 자료 중에서 기준값의 비율로 표현되는 척도
- 절대 0값을 가진다는 의미이기도 함
- 모든 연산이 가능
- 기준값이 단위가 됨
- 예
  - 길이, 중량, 시간, 밀도, 각도, 거리 등

자료의 정리

Namun Cho/ mailto:namun@snu.ac.kr

주가수익률 데이터
P/E ratio (PER) Data

- 주가수익률 Price-to-Earning Ratio
  - 기업의 EPS (earning per share)에 대하여 투자자들이 주식가격으로 얼마나 지불하는가를 나타냄
- 낮은 P/E ratio는 주식이 저평가되었다는 것을 의미하지만 반드시 그렇게 해석되어야만 하는 것은 아님

$$P/E\ ratio := \frac{\text{주식가격}}{EPS}$$

Raw Data

- 2002년 96개 기업 PER data
- 81 63 98 86 83 44 43 58 55 50 29 28 40 36 32 23 21 28 27 26 16 16 20 19 17 13 13 15 15 14 12 12 13 12 12 11 11 12 12 12 11 11 11 11 11 64 58 10 10 10 43 42 93 83 81 28 28 56 50 48 22 21 36 32 30 16 15 28 27 23 13 13 20 18 17 12 12 15 14 14 11 11 12 12 11 10 12 12 12 11 11 11 12 10 10 10
- 아래 링크로 억세스 가능
  - https://docs.google.com/spreadsheets/d/1r3Lbu0x1qWoFzwHCZGIHVdRJZ_Mxwu_MHd2Cr9uk6lA/edit?usp=sharing
- 위 자료를 어떻게 하면 특성을 파악할 수 있을까?

(상대)도수분포
(Relative) Frequency Distribution

- 자료의 구성을 알아볼 수 있음
- 구간별로 해당되는 관측치의 수 (frequency: 도수, 빈도)를 셈
- 상대도수: 전체 자료 수에서 해당 빈도가 차지하는 비율
- 도수 구간은 데이터의 성격이 잘 드러나는 수준에서 결정해야함

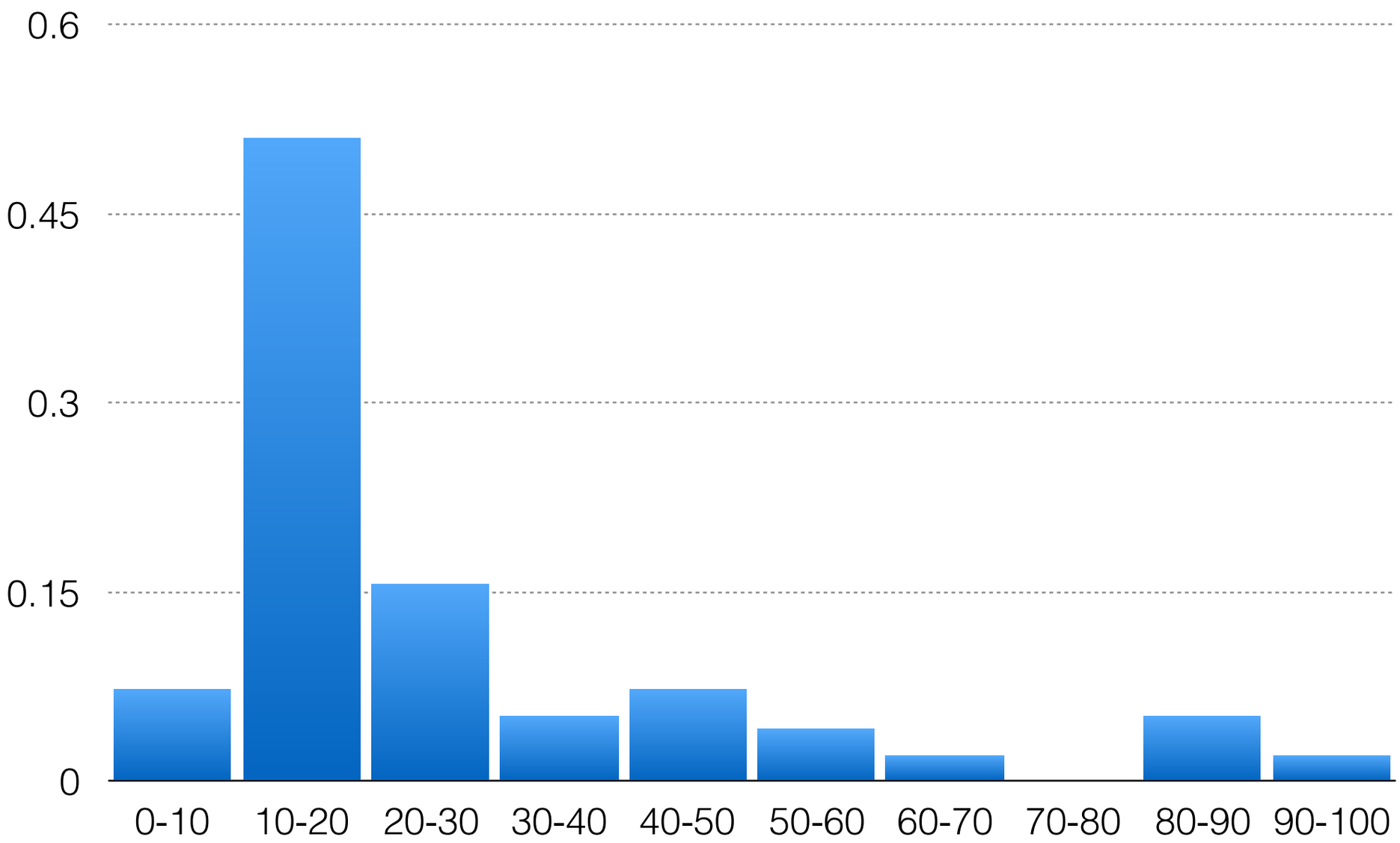
도수분포표
Frequency Distribution Table

Table 1

범위		누적도수	도수	상대도수	%
0	10	96	7	0.0729	7.3
10	20	89	49	0.5104	51
20	30	40	15	0.1563	15.6
30	40	25	5	0.0521	5.2
40	50	20	7	0.0729	7.3
50	60	13	4	0.0417	4.2
60	70	9	2	0.0208	2.1
70	80	7	0	0	0
80	90	7	5	0.0521	5.2
90	100	2	2	0.0208	2.1
	총합		96	1	100

Namun Cho/ mailto:namun@snu.ac.kr

히스토그램



Namun Cho/ mailto:namun@snu.ac.kr

주요 금융 데이터의 이해

보조교재 2장, 4장

Namun Cho/ mailto:namun@snu.ac.kr

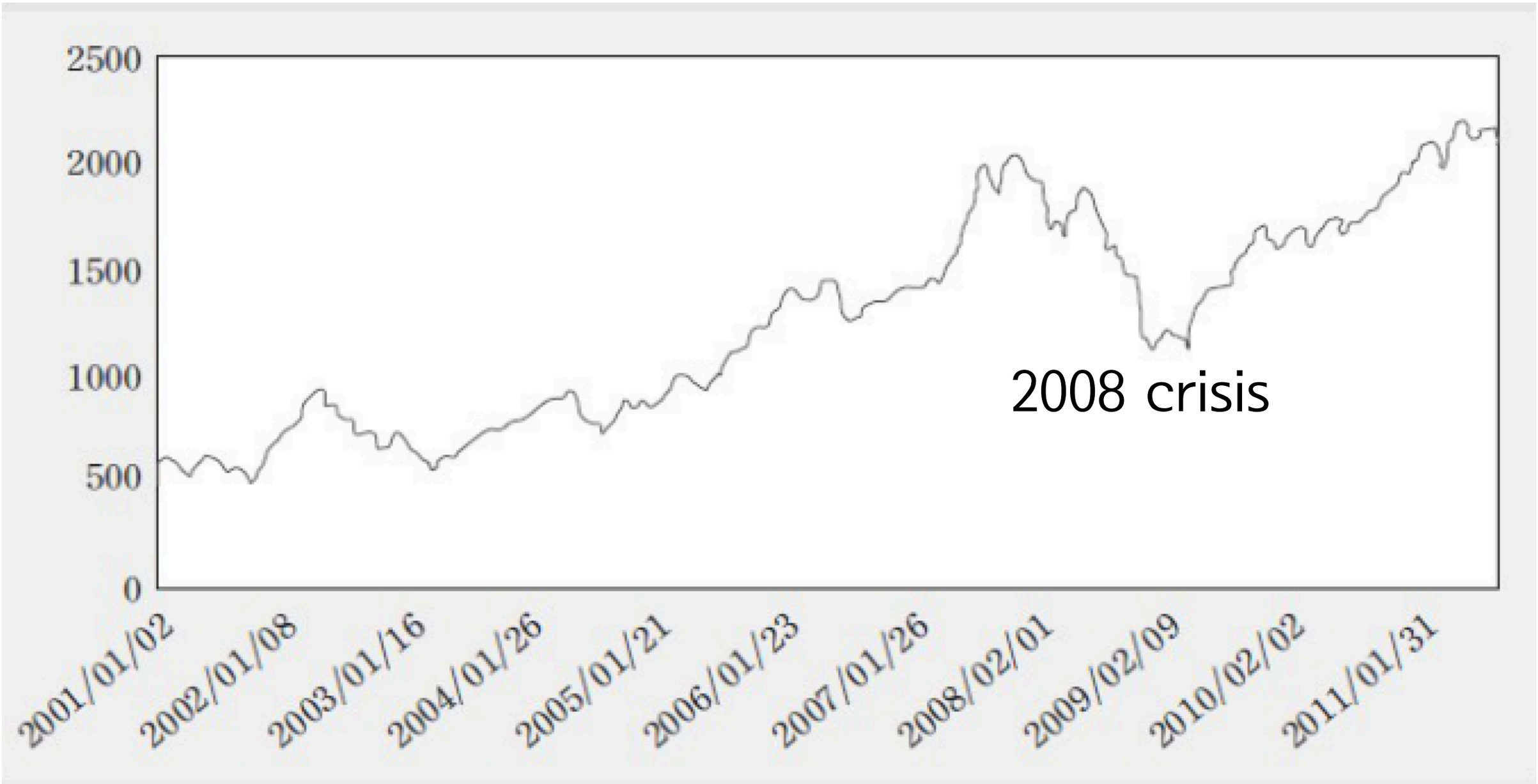
주가지수

- 주식시장 전체의 주식가격 움직임을 대표
  - 최초 주가를 100으로 설정
- 산출법에 따른 분류
  - 시가총액식 주가지수 (가중평균)
    - 모든 주가*상장주식수의 합/기준시점 주가*상장주식수의 합
  - 다우존스식 주가지수 (단순평균)
    - 대표 우량주 주가 합계를 채택종목수로 나눔

한국의 주가지수

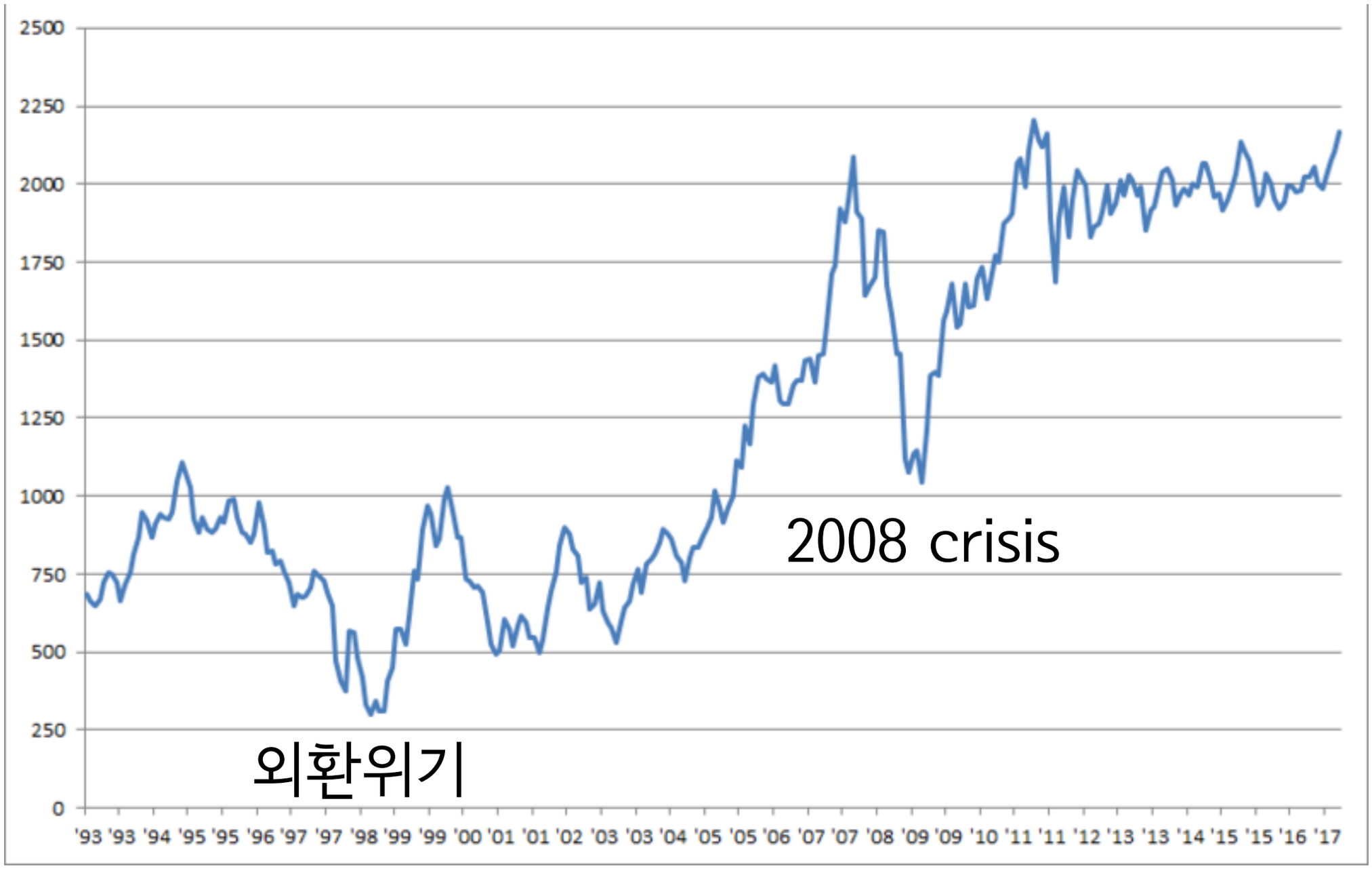
- 한국 종합 주가지수 기준시점: 1980.1.4
  - 이 시점의 지수가 100
- KOSPI 200
  - 대표성이 큰 200기업의 주가지수
- 코스닥 종합지수
  - 상장주식에 대한 시가총액으로 작성

한국 종합주가지수:
2001 - 2011



Namun Cho/ mailto:namun@snu.ac.kr

KOSPI: 1993-2017



Namun Cho/ mailto:namun@snu.ac.kr

주식시세표 (KOSPI)

Namun Cho/ mailto:namun@snu.ac.kr

주식시세표 (KOSPI)

현재 종합주가지수

Namun Cho/ mailto:namun@snu.ac.kr

주식시세표 (KOSPI)

현재 종합주가지수

전일대비 89.43포인트 하락

Namun Cho/ mailto:namun@snu.ac.kr

주식시세표 (KOSPI)

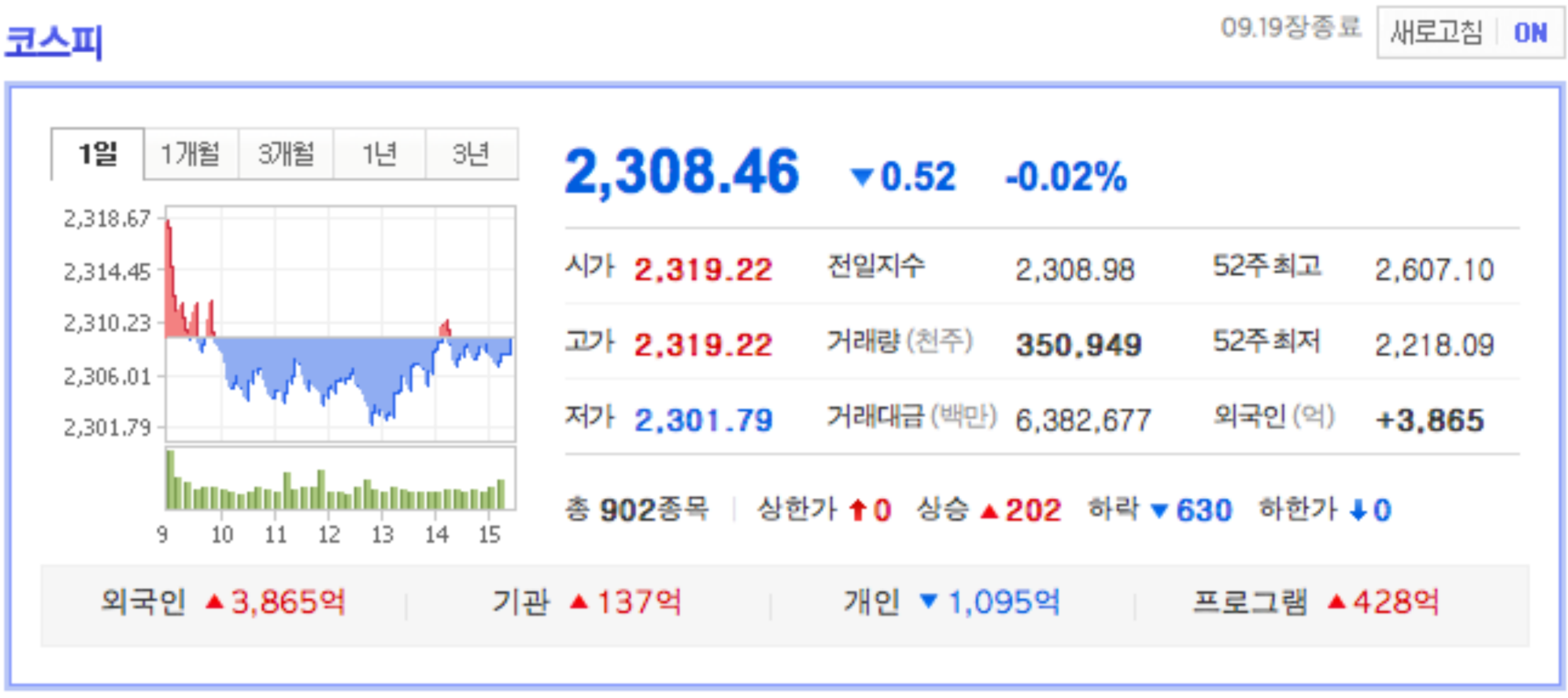
현재 종합주가지수

전일대비 89.43포인트 하락

매입/매도된 주식수

Namun Cho/ mailto:namun@snu.ac.kr

코스피: 2018.9.20



KOSPI: Korea composite Stock Price Index

Namun Cho/ mailto:namun@snu.ac.kr

개별주식시세

건설업 | -2.65%

종목명	현재가	등락률	종목명	현재가	등락률
BS금융지주	14,600	-4.89%	DGB금융지주	16,300	-1.81%
HMC투자증권	17,950	-3.75%	KB금융	47,150	-2.38%

Namun Cho/ mailto:namun@snu.ac.kr

개별주식시세

건설업 | -2.65%

종목명	현재가	등락률	종목명	현재가	등락률
BS금융지주	14,600	-4.89%	DGB금융지주	16,300	-1.81%
HMC투자증권	17,950	-3.75%	KB금융	47,150	-2.38%

현재 주당 가격

Namun Cho/ mailto:namun@snu.ac.kr

개별주식시세

건설업 | -2.65%

종목명	현재가	등락률	종목명	현재가	등락률
BS금융지주	14,600	-4.89%	DGB금융지주	16,300	-1.81%
HMC투자증권	17,950	-3.75%	KB금융	47,150	-2.38%

현재 주당 가격

전일 종가 대비 현재 가격

Namun Cho/ mailto:namun@snu.ac.kr

주가 데이터의 측정주기

- 연도: 1년에 1개 자료
- 분기: 연간 4개
- 월별: 연간 12개
- 일별: 주당 5개 (월-금)
- 분: 1시간에 60개
- 틱: 거래 발생시마다
  - 틱자료는 유상 (코스콤)

주가 자료의 측정값

- 종가로 측정
  - 측정주기에서 가장 마지막 틱에 거래된 값
- 통화량 데이터는 종가에 해당하는 말잔과 함께 평균에 해당하는 평잔이 함께 측정되고 있음
  - 통계학적 측면에서는 종가보다는 평균이 해당 주기의 움직임을 더 잘 대표함

시계열 자료의 표현

- 금융 데이터는 시계열 (time series) 자료임
- 관측 시점, 그리고 관측시점 사이의 간격 (time lag)이 중요한 정보임
- 시간 정보는 하첨자로 표현

$$Y_t : t = 1,2,3,\cdots \qquad\qquad Y_1, Y_2, Y_3, \cdots$$

시계열도표
Time Series Plot

- 시간 경과에 따른 금융데이터를 그래프로 표현
- 가로축: 시간, 세로축: 금융데이터
- 주로 꺾은선그래프를 사용

종합주가지수: 분별

Namun Cho/ mailto:namun@snu.ac.kr

종합주가지수: 일별

Namun Cho/ mailto:namun@snu.ac.kr

종합주가지수: 주별

Namun Cho/ mailto:namun@snu.ac.kr

종합주가지수: 월별

Namun Cho/ mailto:namun@snu.ac.kr

표본자료의 대표치

Namun Cho/ mailto:namun@snu.ac.kr

표본자료의 특성

- 중심
  - 중심화 측도
- 퍼진 정도
  - 산포 측도
- 기타
  - 치우친 정도: 왜도
  - 집중된 정도: 첨도

중심화 측도
Measure of Central Tendency

- 자료의 중심성을 표현하는 좀 더 압축적 형태
- 분포보다는 정보량이 적음
- 주요 중심화 측도
  - 평균 mean
  - 중위수 median
  - 최빈값 mode

평균
Mean (Average)

- 가장 널리 사용하는 중심 측도
- 산술평균 Arithmetic Mean
- 가중평균 Weighted Mean
- 기하평균 Geometric Mean

Data

- 일단, 수치 데이터로 한정하자.
- 데이터의 형태
  - 수치 데이터의 뭉치
    - 값들과, 값들을 구분하는 구분자 (separator)로 구분
      - txt, csv, tsv
    - 좀 더 자동화된 프로그램의 경우 이를 자동화함
      - 통계프로그램 (R, STATA, SPSS, SAS 등)
      - 스프레드쉬트

자료의 형식적 표현
Data: Formal Representation

- 자료 전체는 집합 X 로 표현
- 자료의 각 값은 x_i 로 표현

$$X = \{x_1, x_2, \cdots, x_N\}$$

$$x_i \in X$$

모산술평균
Population Arithmetic Mean, μ

$$\mu := \frac{1}{N}\sum_{i=1}^{N} x_i$$

- 모집단 전체의 산술평균
- 모집단 전체의 값을 알아야 하므로 쉽게 구할 수 있는 값이 아님
- 표본평균을 통해 추정

표본평균
Sample Mean

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 표집으로 구한 표본들의 산술평균
- (주요 조건을 충족한다는 전제하에) 모산술평균의 좋은 추정치

산술평균의 특성

- 모든 자료값을 사용
- (unique) 주어진 자료에 대해 반드시 하나만 존재함
  - 좋은 특성
- 극단값 (extreme value)에 영향을 받음
  - 나쁜 특성

$$\sum_{i=1}^{N}(x_i-\mu)=\sum_{i=1}^{n}(x_i-\bar{X})=0$$

연습

- 자료를 구하기
- 샘플 추출하기
- 모산술평균, 표본평균을 구해보기
- 스프레드쉬트 사용

연습

Index	Data
1	32
2	83
3	82
4	70
5	29
6	36
7	49
8	90
9	21
10	90

Index	Sample Index	Sample Data
1	9	21
2	1	32
3	8	90
4	1	32
5	10	90

Statistics	Value
Population Arithmetic Mean	58.2
Sample Mean	53

Namun Cho/ mailto:namun@snu.ac.kr

가중평균
Weighted Mean

$$\bar{X}_w := \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- 산술평균의 일반화
  - 산술평균은 모든 자료값의 가중치가 같은 가중평균과 동등한 의미임
- 가중치 w_i

예: 주가수익자료

주식수	주가수익률	주식수*주가수익률
7	9	63
8	9	72
6	12	72
7	14	98
1	8	8
9	11	99
6	10	60

Statistics	Value
Weighted Mean	10.727272727

기하평균
Geometric Mean

$$G := \left(\prod_{i=1}^{n} x_i \right)^{1/n} = (x_1 x_2 \cdots x_n)^{1/n}$$

- 지수적 성장을 하는 자료의 경우 산술평균보다 좋은 특성을 보임
- 산술평균의 지수적 표현
  - 기하평균의 정의식에 로그를 취하면 로그데이터의 산술평균임을 확인할 수 있음

산술평균 vs. 기하평균

- 예: 자산 가치가 3년간 1억 → 2억 → 10억 → 20억으로 증가한 경우
- Q: 매년 평균적으로 얼마나 증가했나?
  - 산술평균: (2배 + 5배 + 2배)/3 = 3배
    - 하지만 3년간 3배 성장은 27배 (오차가 큼)
  - 기하평균: (2*5*2)^(1/3) ≈ 2.714
    - 3년간 2.714배 성장할 경우: 19.99배 (오차가 적음)

기하평균의 특성

- 자료의 값이 양수이어야 함
- 산술평균보다 같거나 작음
  - 같은 경우는 모든 자료의 값이 같을 경우뿐임
- 산술평균보다 극단값의 영향을 적게 받음

연습: CJ의 5년간 평균 성장률 구하기

- EPS: Earning Per Share
- 주당 순이익
- 구하고자 하는 것:
  - 5년간 연평균 EPS 성장률
  - Average Annual Compound Growth Rate

사업연도	EPS	Growth Rate (%)
제45기	1078	
제46기	7369	583.58
제47기	4815	-34.66
제48기	2179	-54.75
제49기	1981	-9.09

E_t R_t

자산 수익률
Asset Return

- 단순 총수익률 One Period Simple Gross Return

$$\frac{P_t}{P_{t-1}}$$

- 단순 수익률 Simple Return

$$R_t := \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}$$

- k 시점 단순 수익률 k Period Simple Return

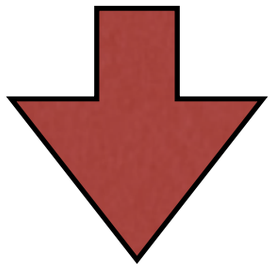
$$1 + R_t(k) = \frac{P_t}{P_{t-k}} = \frac{P_t}{P_{t-1}} \frac{P_{t-1}}{P_{t-2}} \cdots \frac{P_{t-k+1}}{P_{t-k}}$$

$$= (1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1}) = \prod_{j=0}^{k-1} (1 + R_{t-j})$$

연평균 자산수익률

$$1 + R_t(k) = \frac{P_t}{P_{t-k}} = \frac{P_t}{P_{t-1}} \frac{P_{t-1}}{P_{t-2}} \cdots \frac{P_{t-k+1}}{P_{t-k}}$$

$$= (1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1}) = \prod_{j=0}^{k-1} (1 + R_{t-j})$$



$$\text{Annualized}[R_t(k)] = \left[\prod_{j=0}^{k-1} (1 + R_{t-j}) \right]^{1/k} - 1$$

연평균 성장률

$$\text{Annualized}[R_t(5)] = (1981/1078)^{1/5} - 1 \approx 0.164$$

사업연도	EPS	Growth Rate (%)
제45기	1078	
제46기	7369	583.58
제47기	4815	-34.66
제48기	2179	-54.75
제49기	1981	-9.09

E_t R_t

Namun Cho/ mailto:namun@snu.ac.kr

중위수
Median

- 주어진 n개의 자료를 크기순으로 정렬(sort)했을 때 (n+1)/2 번째 수
- 자료의 갯수가 홀수 일때는 단일수
- 자료의 갯수가 짝수 일때는 n/2번째 값과 n/2 + 1 번째 값의 산술평균
- n분위수의 n/2 분위

1, 3, 3, **6**, 7, 8, 9

Median = **6**

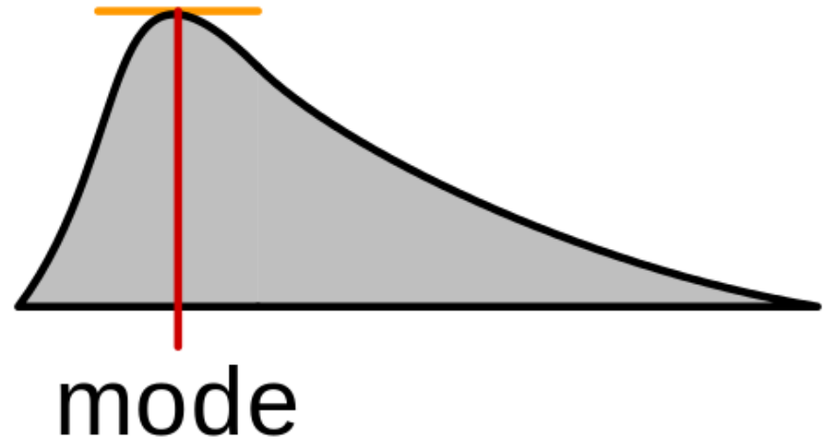
1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

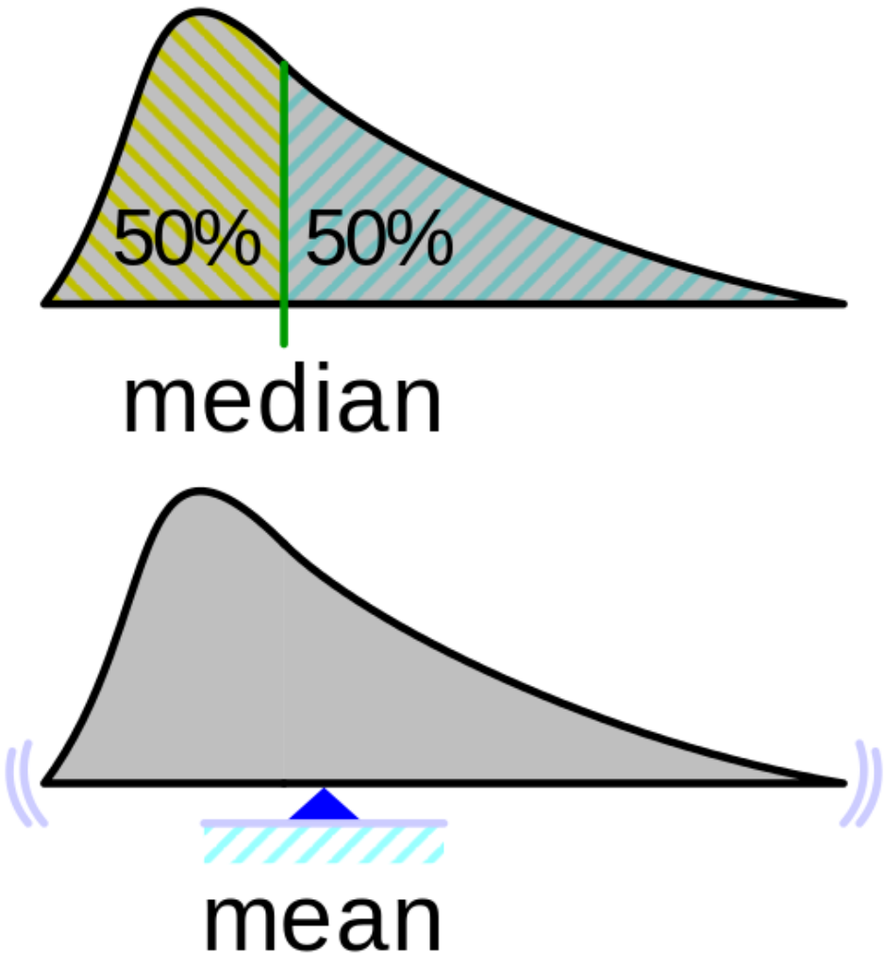
= **4.5**

https://en.wikipedia.org/wiki/Median#/media/File:Finding_the_median.png

최빈값
Mode

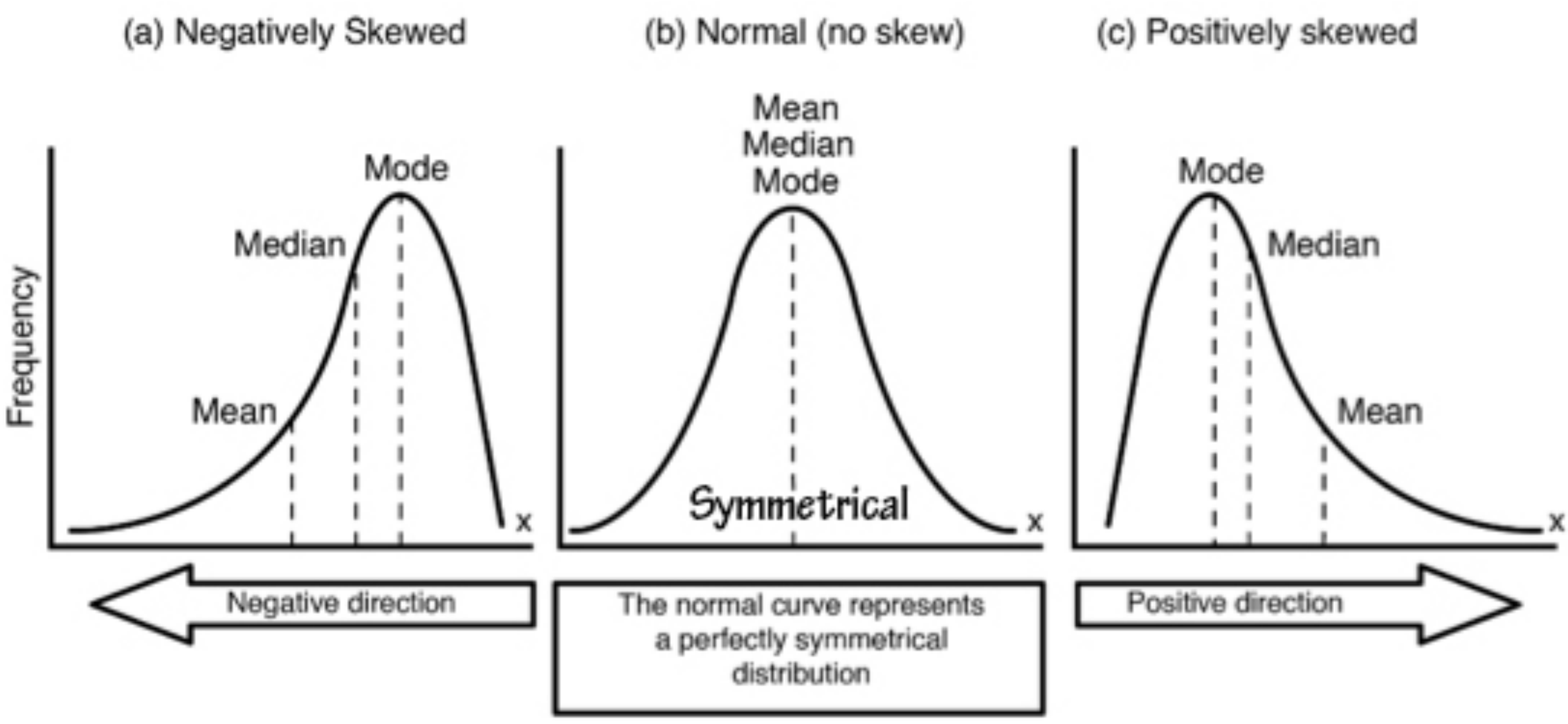


- 빈도수가 최대인 자료값
  - 빈도수: 자료값이 나타난 횟수
- 예:
  - {11,14,13,13,12,50,13}
  - 최빈값: 13
  - 자료 13의 빈도수: 3



https://en.wikipedia.org/wiki/Mode_(statistics)#/media/File:Visualisation_mode_median_mean.svg

분포에 따른 산술평균, 중위수, 최빈값의 관계



Namun Cho/ mailto:namun@snu.ac.kr

산술평균, 중위수, 최빈값

	Mean	Median	Mode
유일성 (Unique)	Yes	Yes	No
모든 자료를 사용	Yes	No	No
극단값 영향	Yes	No	No
자료재정렬 필요성	No	Yes	Yes

보라색 영역: Bad Property

금융 응용 연습

Namun Cho/ mailto:namun@snu.ac.kr

포트폴리오

- 투자금을 여러 종목에 나누어 투자한 내역
  - 위험을 줄이기 위한 목적

포트폴리오의 예:
상호기금 포트폴리오
the stock portfolio of a mutual fund

TABLE 25-1 Fidelity Spartan 500 Index Fund, Top Holdings (as of November 2014)

Company	Percent of mutual fund assets invested in a company
Apple Inc.	3.4%
Exxon Mobil Corp.	2.3
Microsoft Corp.	1.8
S&P 500 Index Future	1.7
Johnson & Johnson	1.6
General Electric Co.	1.4
Berkshire Hathaway Inc.	1.3
Wells Fargo & Co.	1.3
Chevron Corp.	1.3
JPMorgan Chase & Co.	1.2

Source: Fidelity Investments.

포트폴리오의 수익률

- 수익률의 기본원칙
  := [추가로 얻은 금액] / [투자금액]
- 따라서 포트폴리오의 수익률
  = [포트폴리오 투자로 얻은 금액] / [포트폴리오에 투자한 금액]
- 연수익률로 환산하기 위해서는 기하평균을 사용

예제 3.1

- 구매금액 (분모)
  1910*100+750*200
- 수입금
  1850*100+100*8+900*200+
  5*200
- 순수입금 (분자)
  = 수입금 - 구매금액
- 7.5%

	A사 주식	B사 주식
구매가격	1910	750
구매수량	100	200
현재가격(1년후)	1850	900
배당액	8	5

예제 3.2

- 우측표: 어떤 포트폴리오의 연간 수익률
- Q: 만일 이 관리자에게 1999년초에 1000만원을 투자했다면 2002년말에는 얼마가 되어 있을까?

Year	R[t]
1999	10%
2000	7%
2001	-4%
2002	11%

$$1000 \times (1 + 0.1)(1 + 0.07)(1 - 0.04)(1 + 0.11) \approx 1254$$

3.2 포트폴리오의 연평균 수익률

$$[(1+0.1)(1+0.07)(1-0.04)(1+0.11)]^{1/4}-1 \approx 0.058 = 5.8\,\%$$

- 4년간 수익률의 기하평균
- 맞는지 체크하기 위해 1000*(1.058)^4 를 계산해 보고 1254와 비슷한 값이 나오는지 확인하라
- 산술평균으로 계산할 경우
  - (0.1+0.07-0.04+0.11)/4 = 0.0625
  - 1000*(1+0.0625)^4 ≈ 1274 원임

예제 3.3: 불확실한 투자

- 투자금 1000만원
- 40% 수익을 얻을 확률: 1/2 → H(high)라고 명명
- 20% 손실 (-20% 수익)을 얻을 확률: 1/2 → L(low)이라고 명명
- 2년후 기대 수익률은 얼마?

경우의 수
Cases

1년차	2년차	1년차금액	2년차금액	확률	2년차 기대값
H	H	1400	1960	0.25	490
H	L	1400	1120	0.25	280
L	H	800	1120	0.25	280
L	L	800	640	0.25	160
	총합	4400	4840	1	1210

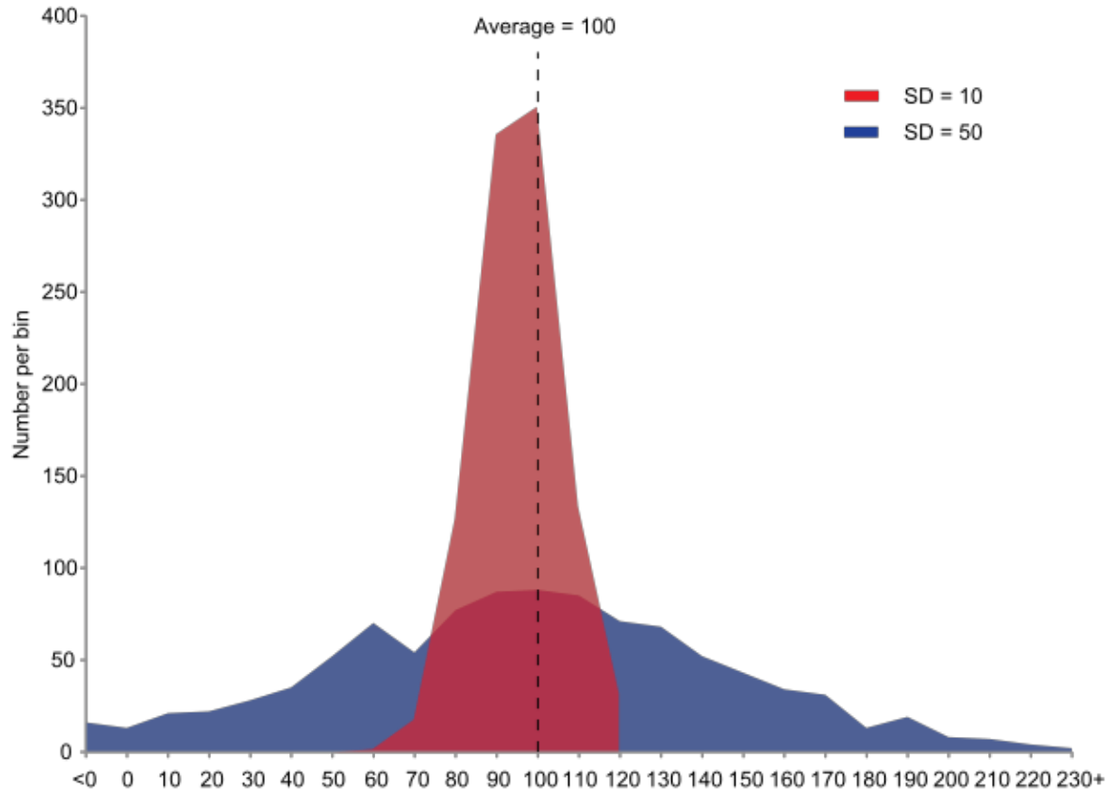
Namun Cho/ mailto:namun@snu.ac.kr

연평균 수익률

- 2년차 기대수익: 1210
- $1210 = 1000(1+r)^2 \Rightarrow r = 0.1 = 10\%$

산포 측도
Measures of Dispersion

- 자료의 퍼진 정도에 대한 통계량
- 퍼진정도:
  - BLUE > RED
- 금융에서는 이를 변동성 (volatility) 이라고 하기도 함
- 예:
  - (3%, -40%, 50%)의 포트폴리오
  - versus (2%, 0%, 3%)의 포트폴리오
  - 후자의 변동성이 더 적음 ⇒ lower risk



https://en.wikipedia.org/wiki/Statistical_dispersion#/media/File:Comparison_standard_deviations.svg

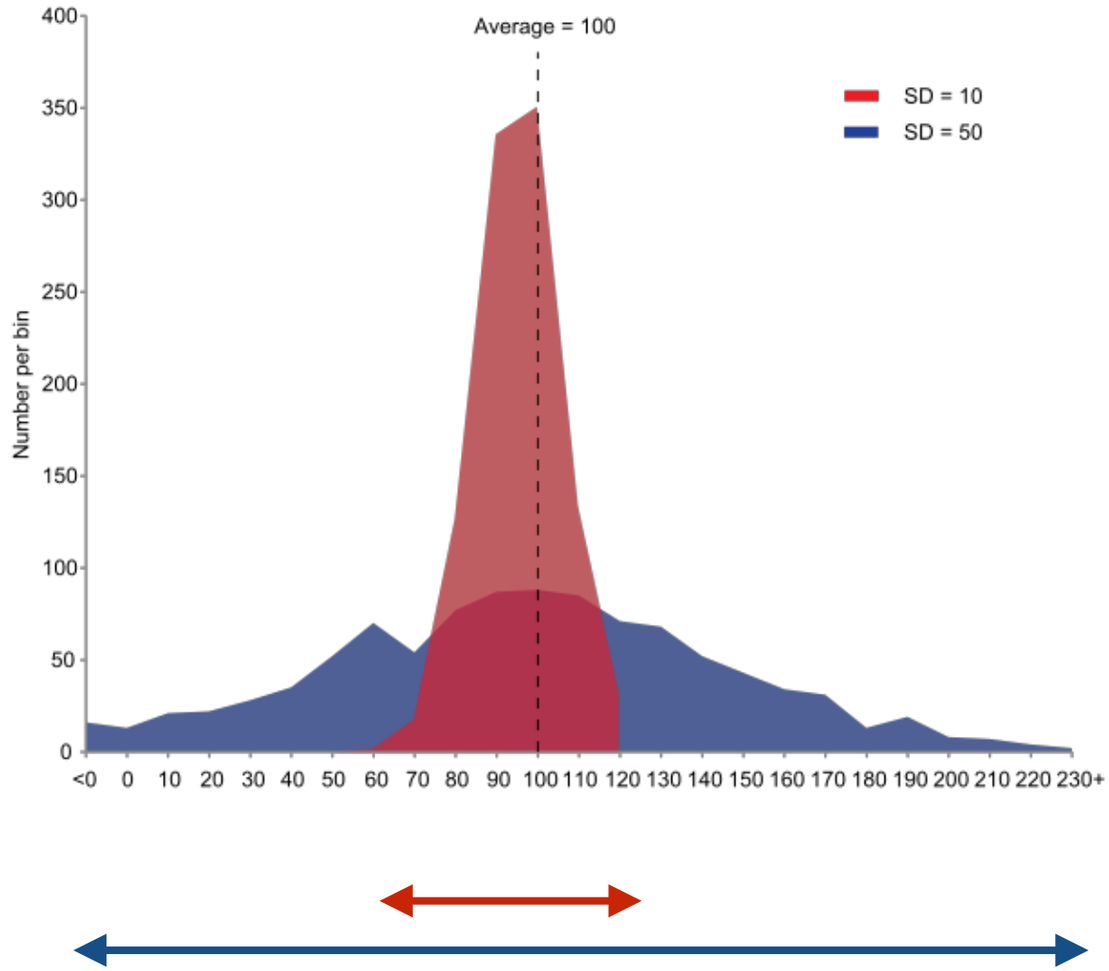
산포 측도의 종류

- 범위 Range
- 절대평균편차 MAD: Mean Absolute Deviation
- 분산 Variance
- 표준편차 Standard Deviation
- 변동계수 Coefficient of Variation
- 샤프지수 Sharpe Ratio

범위
Range

Range := $\max(X) - \min(X)$

- 자료의 최대값과 최소값의 차이
- 장점: 계산이 쉽다
- 단점: 오로지 두 자료의 값만 사용



Namun Cho/ mailto:namun@snu.ac.kr

절대표준편차
Mean Absolute Deviation

- 자료값(xi)과 산술평균(bar X)의 거리의 산술평균
- 계산적 특성상 최소 MAD를 만드는 선형함수는 유일하지 않음
  - 참고: 최소 분산을 만드는 선형함수는 유일함 (OLS)

$$\text{MAD} := \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{X}\right|$$

분산
Variance

- 모분산 Population Variance
- 모표준편차 Population Standard Deviation
- 표본분산 Sample Variance
- 표본표준편차 Sample Standard Deviation

$$\sigma^2 := E(X-\mu)^2$$

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

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

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

분산의 특성

- 평균으로부터 거리가 먼 자료 에 가중치 부과
  - 제곱의 평균이므로
- 수리적으로 좋은 성질 → 산포 측도로 가장 많이 사용
- 표준편차 := 분산^0.5

체비셰프 부등식
Chevyshev's Inequality

$$P(|x-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

- μ 중심으로 k*σ 범위 내에 전체 자료의 100(1-1/k²)% 가 포함됨을 의미
- 모집단 분포와 무관함: 장점

$$P(\mu - k\sigma \leq x \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

변동계수
Coefficient of Variation

- 측정 단위에 따른 문제를 해결하는 산포측도
- 자료 측정단위가 작은 자료들을 비교하려 할 때 CV를 비교

$$CV := \frac{\sigma}{\mu}$$

샤프지수
Sharpe Ratio

- Rp: 포트폴리오 수익률
- RF: 무위험 자산 (Risk-free Asset)의 수익률
    - 국공채 등
- σp : 포트폴리오 수익률의 표준편차
- 의미: 위험자산 투자로 얻는 초과이득
- 높을 수록 유리
- 리스크 프리미엄의 일종

$$\text{Sharpe Ratio} := \frac{\overline{R}_p - \overline{R}_F}{\sigma_p}$$

그룹화 자료로부터 통계량 계산하기

- 중심값 mid point
  - 구간의 상한(UM)과 하한(LM)의 산술평균
  - (UM+LM)/2
- 중심값으로 간주하고 근사치 계산

구간	도수	누적도수	중심값
-25<x≤-20	1	1	-22.5
-20<x≤-15	0	1	-17.5
-15<x≤-10	3	4	-12.5
…	8	12	-7.5
	10	22	-2.5
	20	42	2.5
	8	50	7.5
	4	54	12.5
15<x≤20	2	56	17.5

<표3.7: 디즈니사 주식의 월간로그수익률 자료의 도수분포표>

그 밖의 통계량

- 왜도
  - 분포의 비대칭정도
- 첨도
  - 분포가 집중된 정도
- 분위수
  - 분포의 요약

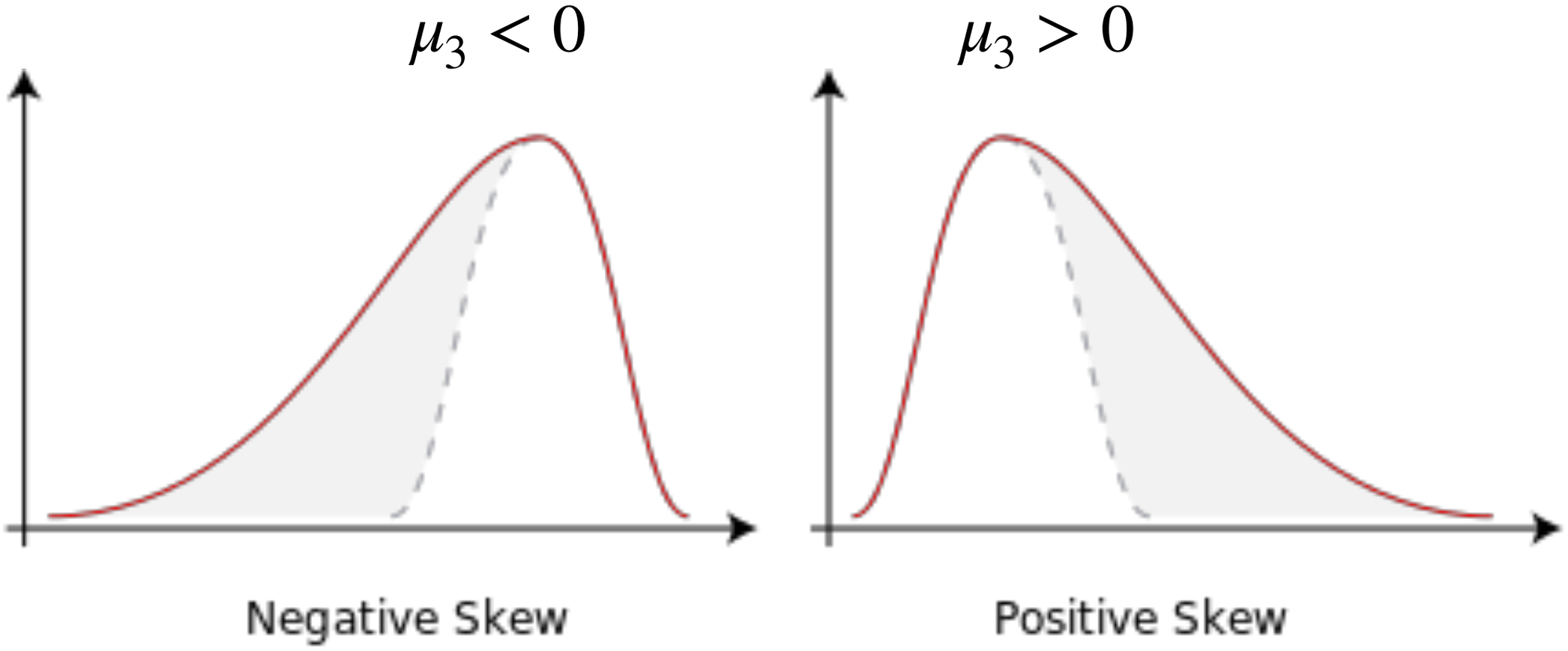
왜도계수: 정의
Coefficient of Skewness

$$\mu_3 := \frac{E(X-\mu)^3}{\sigma^3}$$

- 모왜도계수
- 표본왜도계수

$$\hat{\mu}_3 := \frac{n}{(n-1)(n-2)}\left(\frac{\sum_i^n (x_i-\bar{x})^3}{s^3}\right)$$

왜도계수와 비대칭성



첨도계수: 정의
Coefficient of Kurtosis

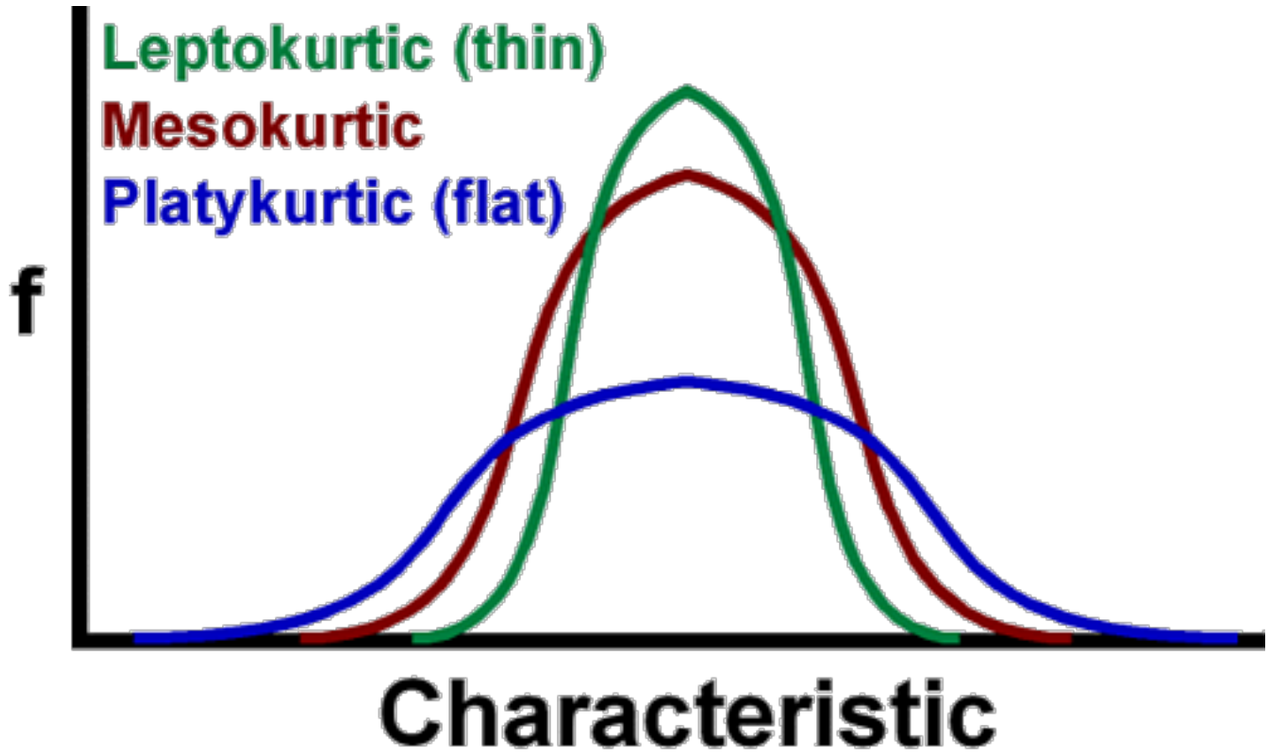
$$\mu_4 := \frac{E(X-\mu)^4}{\sigma^4}$$

- 모첨도계수
- 표본첨도계수

$$\hat{\mu}_4 := \frac{n(n-1)}{(n-1)(n-2)(n-3)} \left(\frac{\sum_i^n (x_i - \overline{x})^4}{s^4} \right)$$

첨도계수와 자료분포

- 첨도의 기준: 정규분포 (첨도계수 = 3)
  - 0 기준으로 +/- 평가를 위해 첨도계수를 μ4-3 으로 정의내리기도 함
  - mesokurtic
- μ4>3 : 긴꼬리분포 leptokurtic
- μ4<3 : 짧은꼬리분포 platykurtic



https://www.quora.com/How-can-I-understand-different-types-of-kurtosis

m 분위수

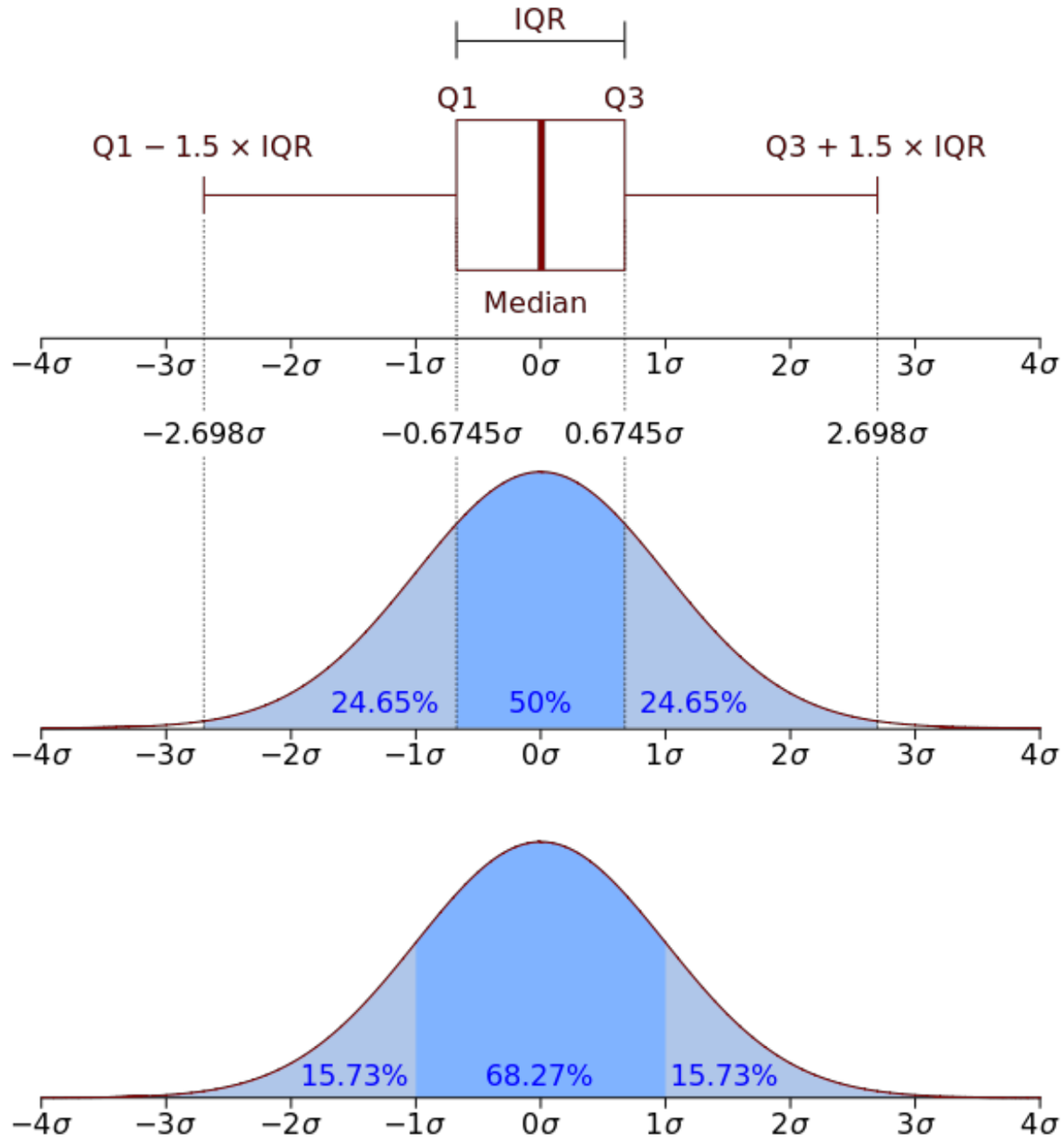
- 4분위수 quartiles
- 5분위수 quintiles
- 10분위수 deciles
- 100분위수 percentiles

- 계산절차
  - 자료를 정렬한다
    - 작은 자료 ~ 큰 자료
  - 분위수를 구한다

$$\text{m분위수} := \frac{y}{m}(n+1), \quad y = 1,2,\cdots,m-1$$

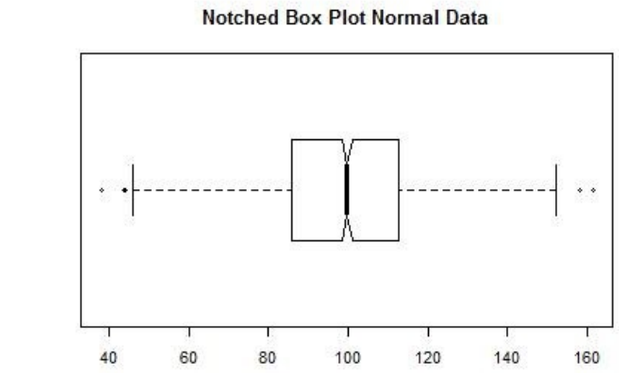
수염상자그림
Box and Whisker Plot

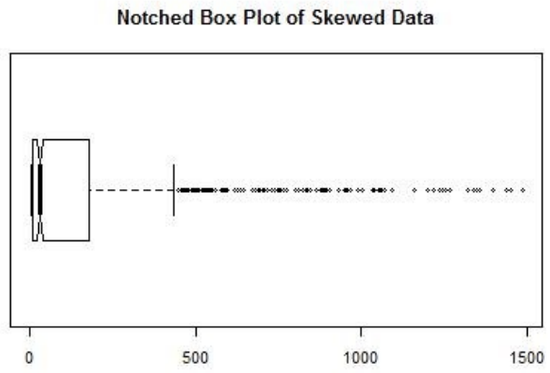
- 100분위수 중
  - Q25 (25 percentile)
    - ≡ Q1 (1 quartile)
  - Q75 (75 percentile)
    - ≡ Q3 (3 quartile)
    - Q50 = median
- 수염은 ±1.5IQR
  - IQR: InterQuartile Range
  - max는 Q3+1.5IQR보다 작은 값 중 가장 큰 값
  - min은 Q1-1.5IQR보다 큰 값 중 가장 작은 값
- 수염 초과값은 Outlier (특이치)

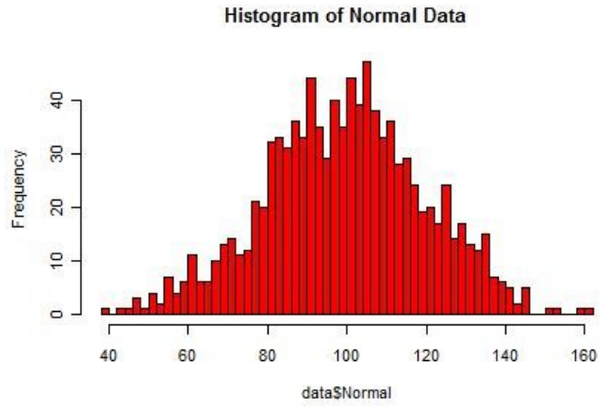


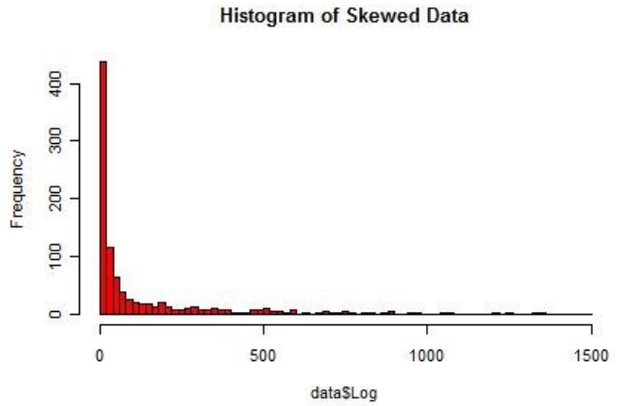
https://en.wikipedia.org/wiki/Box_plot#/media/File:Boxplot_vs_PDF.svg

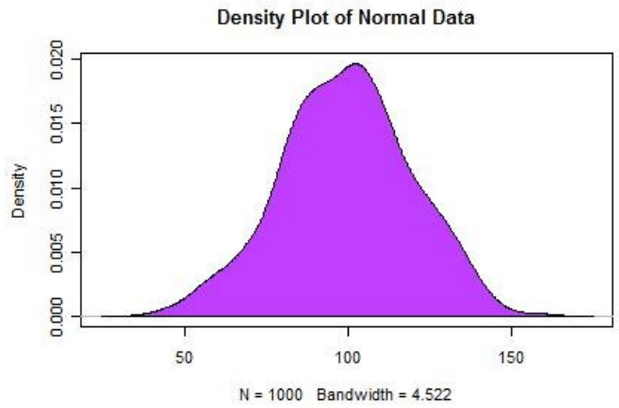
상자그림과 분포

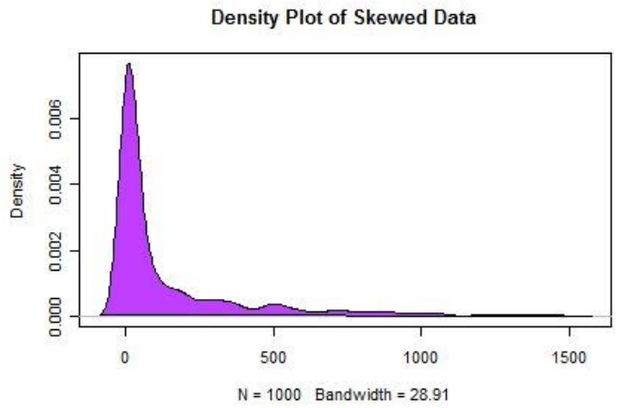












https://sites.google.com/site/davidsstatistics/home/notched-box-plots

Namun Cho/ mailto:namun@snu.ac.kr

과제#1

- 주식 자료 사이트에서 주가 자료 100개 구하기
  - 업종번호는 국내외 주식 자신의 학번 끝 3자리로 검색할 것 (없을 경우 +1 해나가면서 나올 때까지 검색)
  - 일별 종가만 구할것 (과제수행일로부터 과거 100일간)
- 다음의 통계량을 구할 것
  - 산술평균, 중위값, (존재한다면)최빈값, 기하평균, 분산, 표준편차, 변동계수, 왜도계수, 첨도계수, 4분위수
  - 가중평균은 거래주식수로 할 것 (당일거래량/당일주가)
- 본 슬라이드 설명 후 lms 의 과제란에 공시예정
  - 과제란에 스프레드쉬트 파일 포맷으로 업로드할 것

Next Topic

- Ch4: 확률, 확률변수, 기대치

수고하셨습니다!

Namun Cho/ mailto:namun@snu.ac.kr

수고하셨습니다!

Namun Cho/ mailto:namun@snu.ac.kr