

금융 통계학 기본(2) - 기술 통계



이승준 fb.com/plusjune

데이터의 종류

- 질적 데이터 *qualitative data*
- 양적 데이터 *quantitative data*

(한마디로, 수치로 표현되는냐 아니냐로 구분)

질적 데이터 qualitative data (=범주형 categorical)

- 수치로 표현되지 않는 데이터,
- 예를 들어: 전화번호, 성별, 혈액형, 종교 분류, 순위 *rank*, 등급(A,B,C) 등
- 명목 *nominal*: 순서없이 범주 예) 성별, 지역
- 순서 *ordinal*: 순서, 상대적비교 예) 크기, 계층

양적 데이터 quantitative data (=수치적 numerical)

- 수치로 측정이 가능한 데이터 (계량 데이터)
- 예를 들어 : 온도, 가격, 수익률, 주가지수, 실업률, 매출액, 임직원의 수 등
- 연속 continuous: 예) 키, 몸무게, 온도 (실수로 표현)
- 이산 discrete: 예) 몇개의 값, 회수, 가족수 (정수로 표현)

척도

척도^{scale}: 관찰된 결과에 특정 값을 할당하기 위해 사용되는 측정 수준

척도	핵심	특징	예
명목척도 <small>nominal scale</small>	어떤 범주에 속하는가	순서나 크기의 의미 없음	종교, 인종, 성별, 지지정당
순서척도 <small>ordinal scale</small>	순위 부여	등간격 아님, 연산 불가	5점 척도 만족도
구간척도 <small>interval scale</small>	명목/서열 척도의 특성 + 등간격	크기비교(차이) 의미 있음	섭씨온도, 물가지수, 주가지수
비율척도 <small>ratio scale</small>	구간척도의 특성 + 절대 원점	크기비교(차이)와 비율도 의미	월평균 소득, 가족수, 수익률

- 순서척도는 범주형 데이터
- 정보의 양 : 명목척도 < 순서척도 < 구간척도 < 비율척도

질적(=범주형) 데이터 그래픽 요약

- 도수분포표 frequency table
- 막대차트 bar chart
- 파이차트 pie chart

양적(=수치적) 데이터 그래픽 요약

- 점 그래프 dot plot
- 줄기 잎 그래프 stem plot
- 박스 그래프 box plot
- 선 그래프 line plot
- 히스토그램 histogram
- 시계열 그래프 time series graph

데이터의 수치 요약 방법

- ① 중심경향치 measure of central tendency (=대푯값)
- ② 산포도 measure of dispersion
- ③ 상대적 위치의 측도

① 중심경향치(=대푯값)

모집단의 대표적인 경향을 중심경향 *central tendency* 이라고 하며,
중심경향을 나타내는 값을 중심경향치 혹은 대푯값 *representative value* 이라고 한다

중심경향치로 주로 평균값, 중앙값, 최빈값이 사용된다.

- 평균값 *mean*
- 중앙값 *median*
- 최빈값 *mode*

② 산포도

산포(퍼진 *dispersion*), 도(정도 *measure*) (~~먹는 과일 포도와 아무런 관련 없음~~)

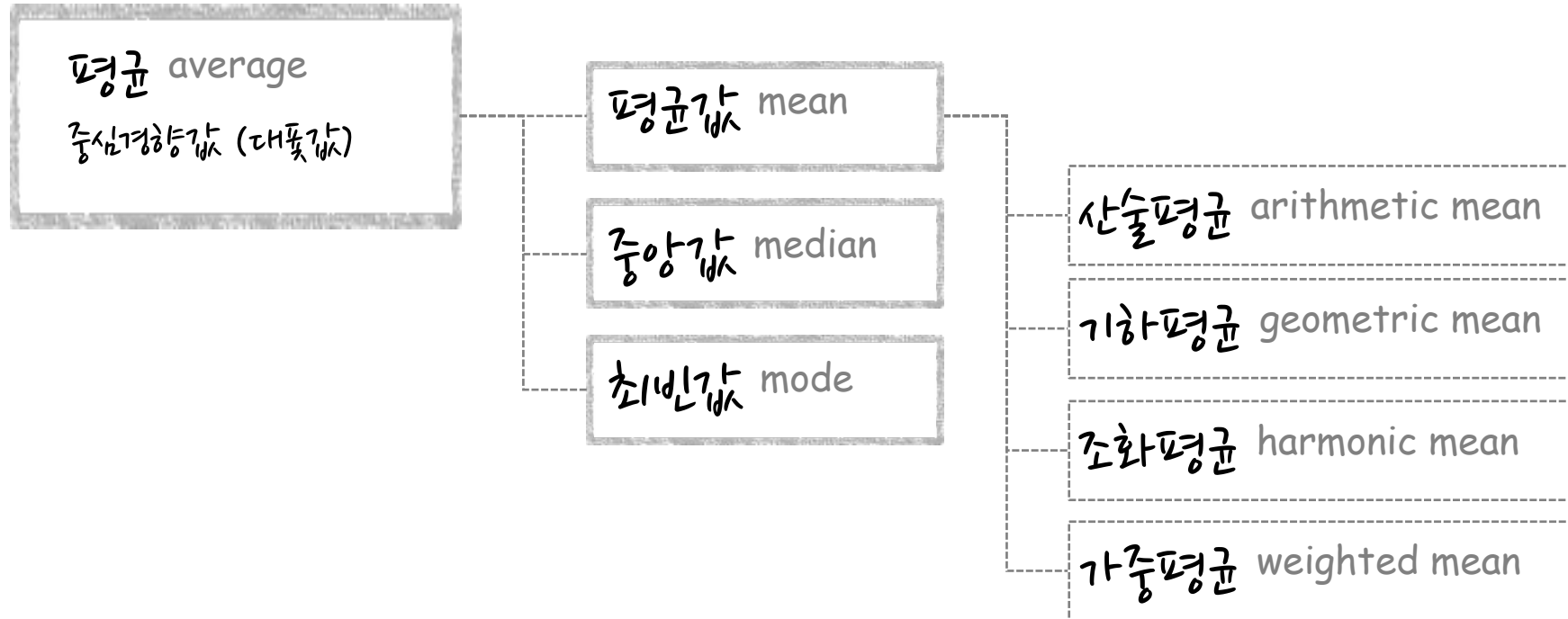
- 분산과 표준편차
- 범위 *range*: 최대 - 최소
- 사분범위 IQR *interquartile range*
- 변동계수 *coefficient of variation*: $CV = \text{표준편차} / \text{산술평균}$

③ 상대적 위치의 측도

특정 값이 전체 데이터에서 어느 정도 위치에 있는지 찾기 위해 사용

- 백분위수: =PERCENTILE(array, k)
- 백분율 순위: =PERCENTRANK(array, x, significance)
- 표준점수: =STANDARDIZE(x, mean, standard_dev)

평균과 평균값



흔히, 평균 average 과 산술평균 arithmetic mean 을 혼용 (~~그러나, 먹고 사는데 지장은 없다~~)

산술평균 arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + \cdots + x_n)$$

자른 전체의 합을
개수로 나눈 값

- 우리가 흔히 "평균"이라고 이야기하는, 데이터 집합의 특징 측정하는 가장 대표 방법
- 예) 특정 주식의 종목 그룹의 가격들의 평균

기하평균 geometric mean

$$G = \sqrt[n]{(x_1 x_2 \cdots x_n)}$$

n개의 양수 값을
모두 곱한 것의 n제곱근

- 누적된 비율, 비례의 평균을 구할 때 (년평균 이익률, 년평균 성장률)
- 첫 해 -50%, 둘째 해 100% 수익이 났다면, 2년간 평균수익률은? 25%일까?
<https://goo.gl/Kd37Wc> [구글다스] 년평균 수익률
- 매출이 작년에 1.5배로 증가하고 올해 6배로 증가했다면?
 $\sqrt[2]{(1.5 \times 6)} = \text{년평균 3배}$

조화평균 harmonic mean

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

주어진 데이터의 역수들을
산술 평균한 값의 역수

- 일정한 값(거리)에 대해 자른의 값(속도)이 바뀌는 때 (평균속도, 작업효율)
- 예) 같은 거리를 갈 때는 8km/h, $\frac{4}{3}$ 때는 12km/h였다면 평균속력은? 9.6km/h

가중평균 weighted mean

$$W = \frac{w_1x_1 + w_2x_1 + \dots + w_nx_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum w_i x_i}{\sum w_i}$$

데이터 × 가중치의 합을
가중치의 합으로 나눈 값

- 각 항목의 비중(가중치)을 반영하고자 할 때(포트폴리오 수익률, 시총반영 주가지수)
- 예) 3개의 종목 A, B, C(각각 700만원, 200만원, 100만원) 포트폴리오를 구성했다.
각 종목의 수익률이 20%, 15%, 50%라면 포트폴리오의 평균 수익률은?

<https://goo.gl/oyh5Vd> [구글다스] 가중평균, 포트폴리오 수익

- 엑셀에서는 =SUMPRODUCT(weight, data)/SUM(weight)

평균, 가중 평균 (파이썬)

```
s = np.random.randint(0, 100, 10)
```

```
print (np.average(s))
```

```
print (np.average(s, weights=range(1,11)))
```

기하평균, 조화평균

```
from scipy import stats
```

```
s = np.random.randint(0, 100, 10)
```

```
print (stats.hmean(s))
```

```
print (stats.gmean(s))
```

금융 데이터와 평균값

평균값	핵심	사용예	스프레드시트 함수
산술평균	덧셈 연산, 총합	1인당 GDP	=average()
기하평균	곱셈 연산, 비율	평균 수익률, 연평균 경제성장률, 물가상승률, 인구증가율	=geomean()
조화평균	정해진 범위 데이터가 변화	평균시속	=harmean()
가중평균	항목들이 비중이 다른데이터 집합	포트폴리오의 기대수익률, 소비자 물가지수, 폭탄주 도수	=sumproduct (w, d) /sum(w)

```
from scipy import stats  
import numpy as np
```

```
s = np.random.randint(0, 100, 10)
```

```
# 평균, 이동평균
```

```
print (np.average(s))
```

```
print (np.average(s, weights=range(1,11)))
```

```
# 조화평균, 기하평균
```

```
print (stats.hmean(s))
```

```
print (stats.gmean(s))
```

수익률

- 투자수익률 Returns of Investment, 투자한 자본에 대한 수익(혹은 손실)의 비율

$$\text{수익률} = \frac{(\text{나중가격} - \text{처음가격})}{\text{처음가격}}$$

- 예를 들어, 100원을 투자해서 110원이 되었다면 수익률은 +10%

가격과 수익률

- 가격 *price*: 추세를 가지고 있어 통계적 특성을 분석하기 어렵다
- 수익률 *returns*: 추세가 없고, 시간에 따른 확률 분포의 변화가 작다

※ 금융 분석에서는 가격을 쓰지 않고 수익률을 쓴다

+10%, -10%의 결과는?

하루는 10% 이익을, 하루는 10% 손해를 봤다면 수익율이 0% 일까?

- 1000원 \rightarrow +10% (잔액 1100원) \rightarrow -10% (잔액 990원)

수익과 손실의 순서를 바꾸면?

- 1000원 \rightarrow -10% (잔액 900원) \rightarrow +10% (잔액 990원)

평균 수익률

	년말 가격	수익률	수익률+1
초기값	1000		
1년차	500	-50%	0.5
2년차	1000	100%	2.0

산술평균은 $(-0.5+1)/2 = 0.25$ (25%)

기하평균은 $((-0.5+1)(1+1))^{(1/2)} - 1 = 0\%$

- 1,000원 짜리 주식
보유 첫해 반토박(-50% 수익률), 다음해 두배(100% 수익률)로 상승했다. 수익률은?
- 증가배율을 계산할때는 기하평균으로 계산해야 한다. 결과적으로 0% !

<https://goo.gl/hAKumc> [구글닥스] 산술평균, 기하평균 수익률

일반수익률과 로그수익률

$$\cdot \text{일반 수익률}(R) = \frac{\text{나중가액} - \text{처음가액}}{\text{처음가액}} = \frac{P - P_0}{P_0} = \frac{P}{P_0} - 1$$

$$\cdot \text{로그 수익률}(R) = \ln \frac{\text{나중가액}}{\text{처음가액}} = \ln \frac{P}{P_0} = \ln(P) - \ln(P_0)$$

일반수익률과 로그수익률

거래일	가격	일반수익률	로그수익률
1일	1,000	.	.
2일	1,300	30.00%	26.24%
3일	800	-38.46%	-48.55%
4일	1,300	62.50%	48.55%
5일	1,100	-15.38%	-16.71%
수익률 합계	39%	38.65%	9.53%
최종수익률	10%	10.00%	9.53%

<https://goo.gl/lfSPjZ> [구글다크스] 일반 수익률, 로그 수익률

일반 수익률

- 재투자 해서 발생하는 손실까지 포함
- 각 거래에서 발생하는 손실률의 합과 최종 손실률이 달라진다

로그 수익률

- 최종 수익률과 수익률의 합계가 일치
- 금융 분야에서는 주로 로그 수익률을 사용한다

거래비용

거래를 위해 수반되는 모든 비용

- 수익률 계산이나 시뮬레이션에서는 가격만으로 계산, 거래비용은 따로 고려
- 매매를 자주 할 때는 매우 중요한 요소 (투자전략 시뮬레이션 때는 거래비용을 고려)
- 증권사 수수료(0.015%) 대비 증권거래세(0.3%) 20배, 증권거래세는 모든 증권사 동일

매매 수수료와 세금

어떤 주식을 10만원에 10주를 매수, 수익률이 10% 라고 가정
(매매 수수료 0.015%, 증권거래세 0.3% 가정)

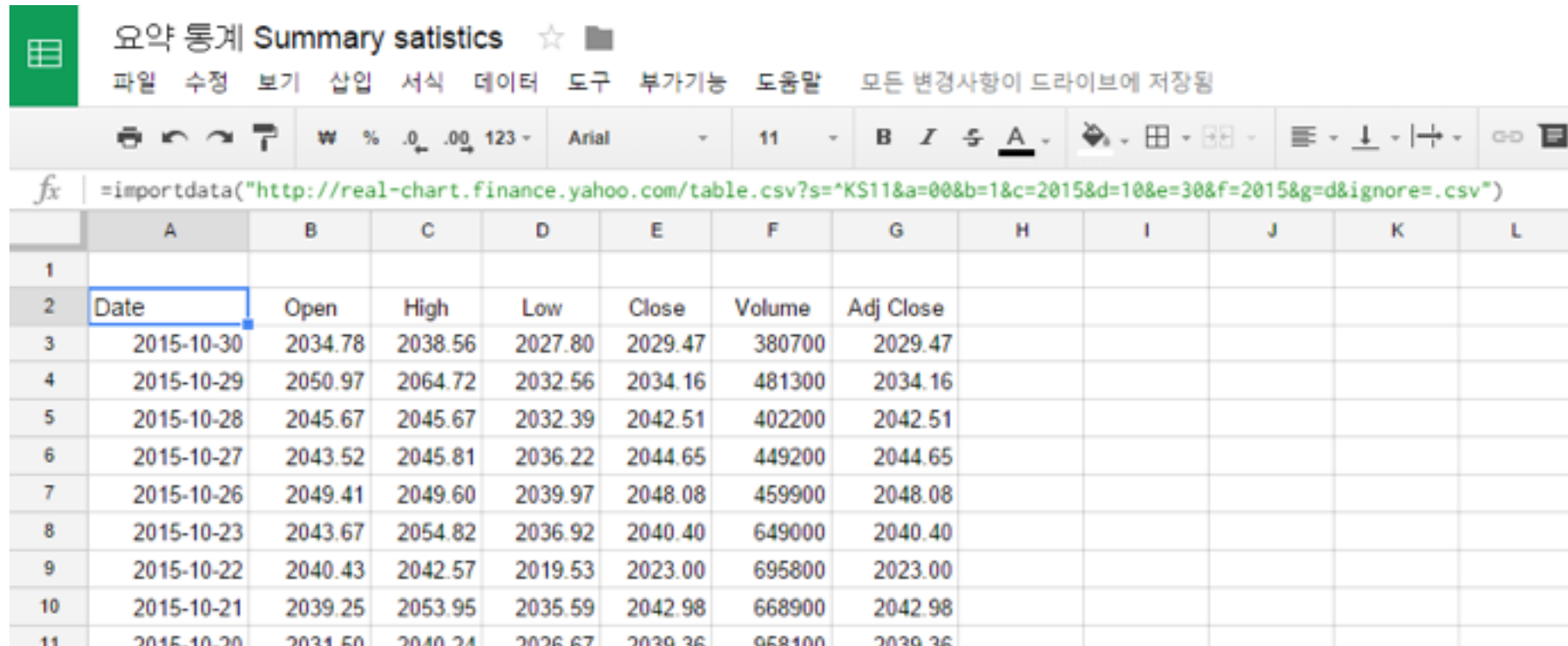
- 총 거래비용은 150원 + 165원 + 3,300원 = 3,615원
 - 매수 수수료: 백만원 x 0.015% = 150원
 - 매도 수수료: 백십만원 x 0.015% = 165원
 - 증권 거래세(매도): 백십만원 x 0.3% = 3,300원

실질수익 = 수익 - (매수수료+매도수수료+세금)

코스피 지수 수익률 (실습)

- 새 시트를 만들고, A2 셀에 다음 입력

```
=importdata("http://real-chart.finance.yahoo.com/table.csv?s=^KS11&a=00&b=1&c=2015&d=10&e=30&f=2015&g=d&ignore=.csv")
```



요약 통계 Summary statistics ☆

파일 수정 보기 삽입 서식 데이터 도구 추가기능 도움말 모든 변경사항이 드라이브에 저장됨

f_x =importdata("http://real-chart.finance.yahoo.com/table.csv?s=^KS11&a=00&b=1&c=2015&d=10&e=30&f=2015&g=d&ignore=.csv")

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	Date	Open	High	Low	Close	Volume	Adj Close					
3	2015-10-30	2034.78	2038.56	2027.80	2029.47	380700	2029.47					
4	2015-10-29	2050.97	2064.72	2032.56	2034.16	481300	2034.16					
5	2015-10-28	2045.67	2045.67	2032.39	2042.51	402200	2042.51					
6	2015-10-27	2043.52	2045.81	2036.22	2044.65	449200	2044.65					
7	2015-10-26	2049.41	2049.60	2039.97	2048.08	459900	2048.08					
8	2015-10-23	2043.67	2054.82	2036.92	2040.40	649000	2040.40					
9	2015-10-22	2040.43	2042.57	2019.53	2023.00	695800	2023.00					
10	2015-10-21	2039.25	2053.95	2035.59	2042.98	668900	2042.98					
11	2015-10-20	2034.50	2040.24	2026.67	2030.36	658100	2030.36					



요약 통계 Summary statistics ☆

파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말 모든 변경사항이 드라이브에 저장됨

fx =ln(G3)-ln(G4) W % .0 .00 123 Arial 10 B I A

	A	B	C	D	E	F	G	H	I
1									
2	Date	Open	High	Low	Close	Volume	Adj Close	Returns	
3	2015-10-30	2034.78	2038.56	2027.80	2029.47	380700	2029.47	-0.00231	
4	2015-10-29	2050.97	2064.72	2032.56	2034.16	481300	2034.16	-0.00410	
5	2015-10-28	2045.67	2045.67	2032.39	2042.51	402200	2042.51	-0.00105	
6	2015-10-27	2043.52	2045.81	2036.22	2044.65	449200	2044.65	-0.00168	
7	2015-10-26	2049.41	2049.60	2039.97	2048.08	459900	2048.08	0.00376	
8	2015-10-23	2043.67	2054.82	2036.92	2040.40	649000	2040.40	0.00856	
9	2015-10-22	2040.43	2042.57	2019.53	2023.00	695800	2023.00	-0.00983	
10	2015-10-21	2039.25	2053.95	2035.59	2042.98	668900	2042.98	0.00177	
11	2015-10-20	2031.50	2040.24	2026.67	2039.36	958100	2039.36	0.00447	
12	2015-10-19	2032.36	2036.30	2022.25	2030.27	726800	2030.27	0.00000	
13	2015-10-16	2037.98	2038.06	2025.19	2030.26	550800	2030.26	-0.00148	
14	2015-10-15	2008.29	2035.80	2007.46	2033.27	498900	2033.27	0.01173	
15	2015-10-14	2014.47	2016.14	2002.63	2009.55	644800	2009.55	-0.00472	
16	2015-10-13	2021.20	2024.50	2010.16	2019.05	605000	2019.05	-0.00128	
17	2015-10-12	2022.33	2030.92	2015.43	2021.63	635100	2021.63	0.00104	
18	2015-10-08	2016.36	2020.19	1998.66	2019.53	701000	2019.53	0.00680	

=ln(G3)-ln(G4)

요약 통계 Summary statistics ☆

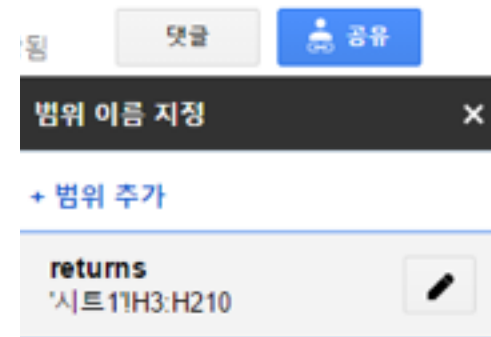
파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말 모든 변경사항이 드라이브에 저장됨

f_x =ln(G3)-ln(G4)

	A	B	C				G	H	I
1									
2	Date	Open	High				Adj Close	Returns	
3	2015-10-30	2034.78	2038.5				2029.47	-0.00231	
4	2015-10-29	2050.97	2064.7				2034.16	-0.00410	
5	2015-10-28	2045.67	2045.6				2042.51	-0.00105	
6	2015-10-27	2043.52	2045.8				2044.65	-0.00168	
7	2015-10-26	2049.41	2049.6				2048.08	0.00376	
8	2015-10-23	2043.67	2054.8				2040.40	0.00856	
9	2015-10-22	2040.43	2042.5				2023.00	-0.00983	
10	2015-10-21	2039.25	2053.9				2042.98	0.00177	
11	2015-10-20	2031.50	2040.2				2039.36	0.00447	
12	2015-10-19	2032.36	2036.3				2030.27	0.00000	
13	2015-10-16	2037.98	2038.0				2030.26	-0.00148	
14	2015-10-15	2008.29	2035.80	2007.46	2033.27	498900	2033.27	0.01173	
15	2015-10-14	2014.47	2016.14	2002.63	2009.55	644800	2009.55	-0.00472	
16	2015-10-13	2021.20	2024.50	2010.16	2019.05	605000	2019.05	-0.00128	
17	2015-10-12	2022.33	2030.92	2015.43	2021.63	635100	2021.63	0.00104	
18	2015-10-08	2016.36	2020.19	1998.66	2019.53	701000	2019.53	0.00680	
19	2015-10-07	1995.87	2006.24	1994.55	2005.84	761700	2005.84	0.00760	
20	2015-10-06	1996.29	1998.02	1985.17	1990.65	661600	1990.65	0.00625	

H열기준 시트 정렬, A → Z
 H열기준 시트 정렬, Z → A
 H열 기준 범위 정렬, A → Z
 H열 기준 범위 정렬, Z → A
 범위 정렬...
 범위 이름 지정...
 보호된 시트와 범위...
 필터
 필터 보기...
 피벗 테이블...
 확인...

수익률 백분율 선택하고
범위 이름을 "returns"로 지정





요약 통계 Summary statistics ☆

파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말 모든 변경사항

fx % .0 .00 123 Arial 11 B I

fx =AVERAGE(returns)

	G	H	I	J	K	L
1						
2	Adj Close	Returns				
3	2029.47	-0.00231				
4	2034.16	-0.00410				
5	2042.51	-0.00105	Mean	평균	0.000250	=AVERAGE(returns)
6	2044.65	-0.00168	Standard Error	표준오차	0.000548	=STDEV(returns)/SQRT(COUNT(returns))
7	2048.08	0.00376	Median	중앙값	0.000163	=MEDIAN(returns)
8	2040.40	0.00856	Mode	최빈값	0.000000	=MODE(returns)
9	2023.00	-0.00983	Standard Deviation	표준편차	0.007899	=STDEV(returns)
10	2042.98	0.00177	Varlance	분산	0.000062	=VAR(returns)
11	2039.36	0.00447	Kurtosis	첨도	1.567933	=KURT(returns)
12	2030.27	0.00000	Skewness	왜도	-0.048146	=SKEW(returns)
13	2030.26	-0.00148	Range	범위	0.054091	=MAX(returns)-MIN(returns)
14	2033.27	0.01173	Minimum	최소값	-0.024967	=MIN(returns)
15	2009.55	-0.00472	Maximum	최대값	0.029124	=MAX(returns)
16	2019.05	-0.00128	Sum	합	0.052101	=SUM(returns)
17	2021.63	0.00104	Count	관측수	208	=COUNT(returns)
18	2019.53	0.00680				
19	2005.84	0.00760				
20	1990.65	0.00625				
21	1978.25	0.00434				

“returns” 영역 데이터에 대한
통계함수 입력

기본 통계량

- 데이터에 대한 간략한 요약 정보.
평균값, 중앙값, 최빈값, 최소값,
최대값, 범위, 표준편차 등
- 엑셀 데이터 분석, “기술통계법”

The screenshot shows an Excel spreadsheet with a stock price dataset. The dataset includes columns for Date, Open, High, Low, Close, Volume, Adj Close, and Returns. A summary statistics table is visible on the right side of the spreadsheet, showing various statistical measures for the 'Returns' column.

The '기술 통계법' (Technical Statistics) dialog box is open, showing the following settings:

- 입력 범위(I): \$H\$3:\$H\$210
- 데이터 방향: 열(C) (selected)
- 첫째 행 이탤릭표 사용(L): ☐
- 출력 옵션: 출력 범위(O): \$J\$3 (selected)
- 새로운 워크시트(P): ☐
- 새로운 통합 문서(W): ☐
- 요약 통계량(S): ☒
- 평균에 대한 신뢰 수준(N): 95 %
- K번째 큰 값(A): 1
- K번째 작은 값(M): 1

Column1	
평균	0.0002505
표준 오차	0.0005477
중앙값	0.0001633
최빈값	0
표준 편차	0.0078995
분산	6.24E-05
첨도	1.5679331
왜도	-0.048146
범위	0.0540913
최소값	-0.024967
최대값	0.0291243
합	0.052101
관측수	208

[구글박스] <https://goo.gl/H90Uom>

요약 통계 Summary statistics

요약 통계 Summary statistics ☆

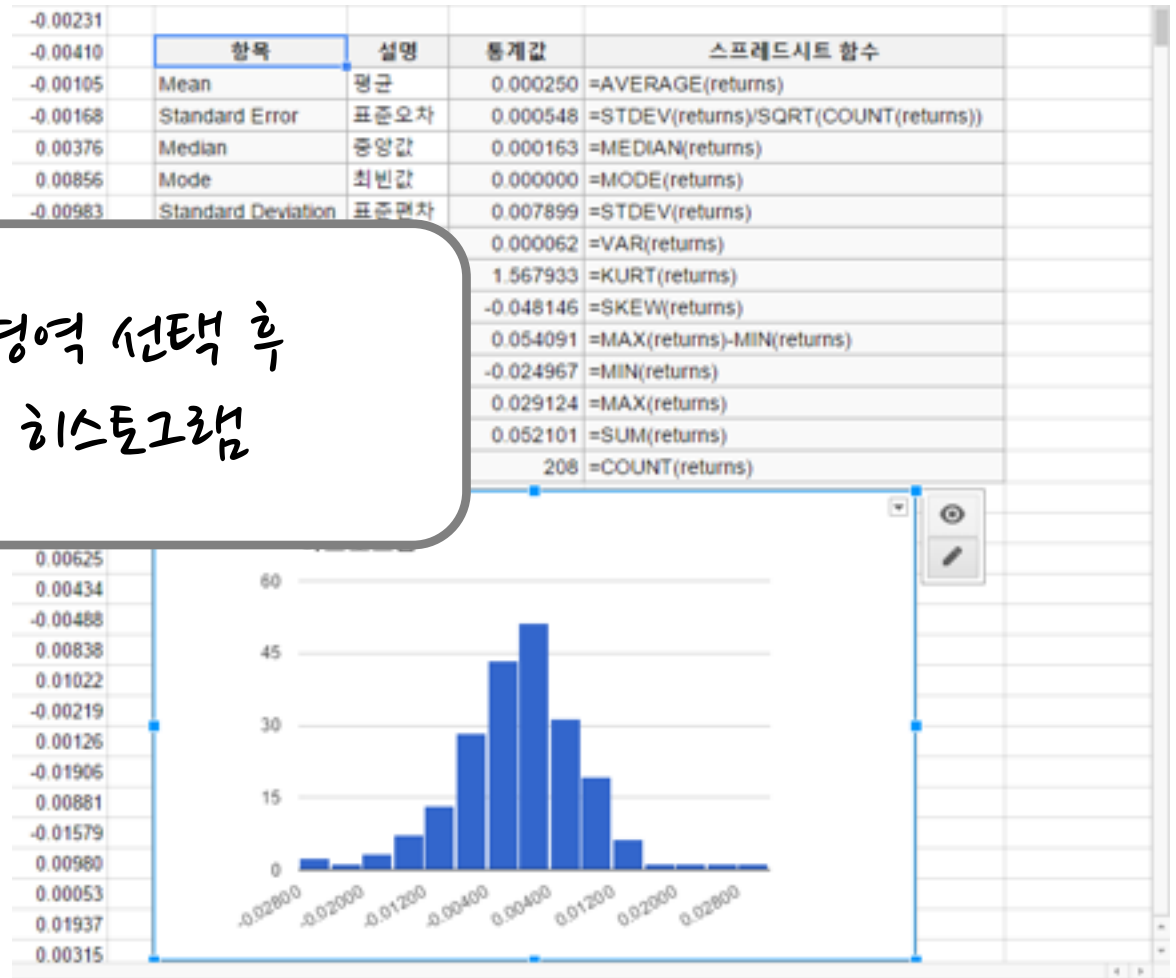
파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말 모든 변경사항이 드라이브

fx =ln(G3)-ln(G4)

위에 행 삽입
아래에 행 삽입
위로 208행 삽입
아래로 208행 삽입
왼쪽에 열 삽입
오른쪽에 열 삽입
새 시트
댓글 Ctrl+Alt+M
메모 Shift+F2
Σ 함수
차트...
이미지...
링크... Ctrl+K
설문지...
그림...

	B	C	F
186	1955.13	1955.52	309800
187	1961.15	1962.15	395700
188	1963.37	1970.27	348000
189	1960.11	1960.35	370000
190	1947.91	1956.10	364800
191	1964.13	1964.13	441500
192	1950.83	1960.12	391000
193	1946.71	1963.53	400900
194	1941.90	1952.41	467000
195	1928.26	1936.40	330600
196	1945.36	1945.36	367700
197	1926.01	1932.50	343100
198	1917.33	1921.88	380000
199	1908.62	1918.85	280900
200	1902.81	1912.42	281100
201	1907.07	1907.07	298700
202	1914.06	1919.69	266500
203	1919.31	1925.68	297500
204	1915.03	1920.77	339000
205	1918.18	1924.66	311900
206	1919.80	1929.10	308800
207	1895.85	1907.23	262200
208	1878.90	1887.60	280500
209	1895.48	1900.02	303700
210	1921.96	1921.96	313400
211	1914.24	1929.15	258800
212			

“returns” 영역 선택 후
차트 삽입, 히스토그램



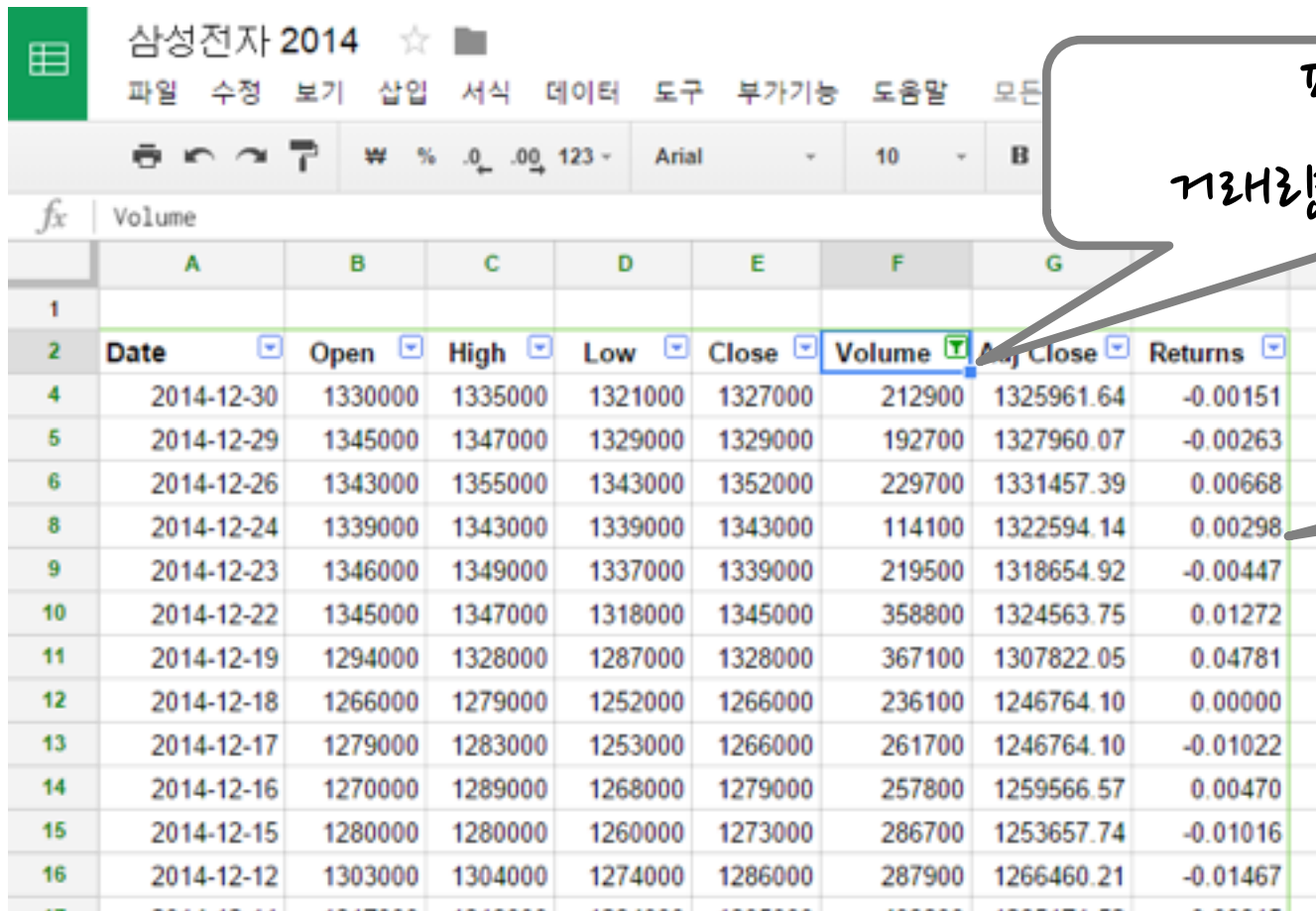
항목	설명	스프레드시트 함수	설명
Mean	평균	=AVERAGE()	자른 전체의 합을 개수로 나눈 값, 산술평균
Standard Error	표준오차	=STDEV()/SQRT(COUNT())	표본의 통계치와 모수와의 차이
Median	중앙값	=MEDIAN()	값의 범위(range)에서 가운데 있는 값
Mode	최빈값	=MODE()	가장 많이 관측되는 값
Standard Deviation	표준편차	=STDEV()	평균에서 떨어진 정도(분산의 제곱근)
Variance	분산	=VAR()	평균에서 떨어진 정도
Kurtosis	첨도	=KURT()	표준(정규분포) 위로 뾰족(>0), 납작(<0)한 정도
Skewness	왜도	=SKEW()	평균으로 부터 왼쪽(+)이나 오른쪽(-)으로 치우친 정도
Range	범위	=MAX()-MIN()	최대값과 최소값의 차이
Minimum	최소값	=MIN()	가장 작은 값
Maximum	최대값	=MAX()	가장 큰 값
Sum	합	=SUM()	모든 관측수의 합계
Count	관측수	=COUNT()	관측수

기본 금융 통계량, 어디에 사용하는가

- 평균, 이동평균: 주가, 기대 수익률, 기댓값, 매매신호
- 왜도, 첨도: 확률 분포의 특성, 시장의 심리 등
- 분산, 표준편차: 투자위험 *risk*, 변동성 *volatility*, 매매신호
- 공분산, 상관계수, 베타: 두 데이터 집합의 비례관계의 정도 (상관분석)

2014년 삼성전자 수익률 (실습)

```
=importdata("http://real-chart.finance.yahoo.com/table.csv?s=005930.KS&a=0&b=1&c=2014&d=11&e=31&f=2014&g=d&ignore=.csv")
```



	A	B	C	D	E	F	G	
2	Date	Open	High	Low	Close	Volume	Adj. Close	Returns
4	2014-12-30	1330000	1335000	1321000	1327000	212900	1325961.64	-0.00151
5	2014-12-29	1345000	1347000	1329000	1329000	192700	1327960.07	-0.00263
6	2014-12-26	1343000	1355000	1343000	1352000	229700	1331457.39	0.00668
8	2014-12-24	1339000	1343000	1339000	1343000	114100	1322594.14	0.00298
9	2014-12-23	1346000	1349000	1337000	1339000	219500	1318654.92	-0.00447
10	2014-12-22	1345000	1347000	1318000	1345000	358800	1324563.75	0.01272
11	2014-12-19	1294000	1328000	1287000	1328000	367100	1307822.05	0.04781
12	2014-12-18	1266000	1279000	1252000	1266000	236100	1246764.10	0.00000
13	2014-12-17	1279000	1283000	1253000	1266000	261700	1246764.10	-0.01022
14	2014-12-16	1270000	1289000	1268000	1279000	257800	1259566.57	0.00470
15	2014-12-15	1280000	1280000	1260000	1273000	286700	1253657.74	-0.01016
16	2014-12-12	1303000	1304000	1274000	1286000	287900	1266460.21	-0.01467

필터링으로
거래량 0인 날 삭제

로그 수익률
 $=\ln(P_0) - \ln(P)$

수익률
히스토그램 추가

종목과 지수 읽기

```
url_tmp = "http://real-chart.finance.yahoo.com/table.csv?" \
          "s=%s&a=0&b=1&c=2014&d=11&e=31&f=2014&g=d&ignore=.csv"
```

```
df_005930 = pd.read_csv(url_tmp % '005930.KS', index_col='Date', parse_dates={'Date'})
```

```
df_005930 = df_005930.ix[df_005930.Volume > 0]
```

```
df_ks11 = pd.read_csv(url_tmp % '^KS11', index_col='Date', parse_dates={'Date'})
```

```
df_ks11 = df_ks11.drop(pd.Timestamp('2014-12-31'))
```

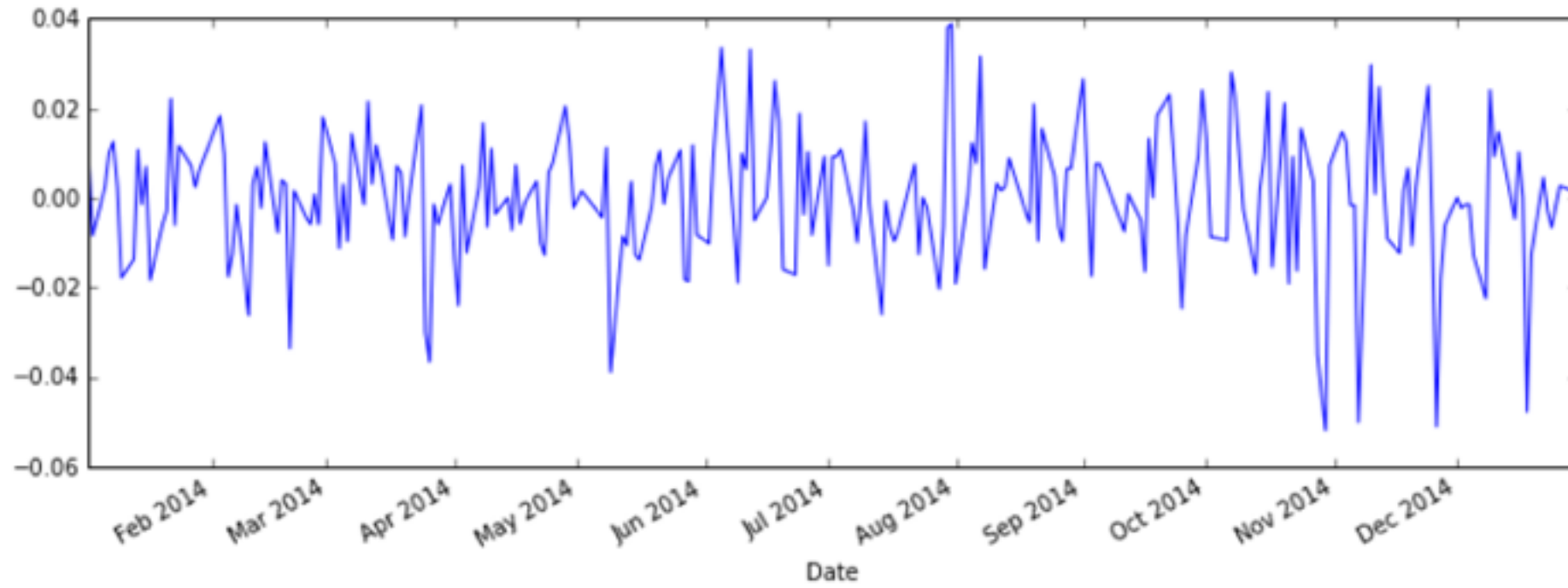

로그 수익률

```
df_005930['Ret'] = np.log(df_005930['Adj Close'] / df_005930['Adj Close'].shift(1))
```

```
df_ks11['Ret'] = np.log(df_ks11['Adj Close'] / df_ks11['Adj Close'].shift(1))
```

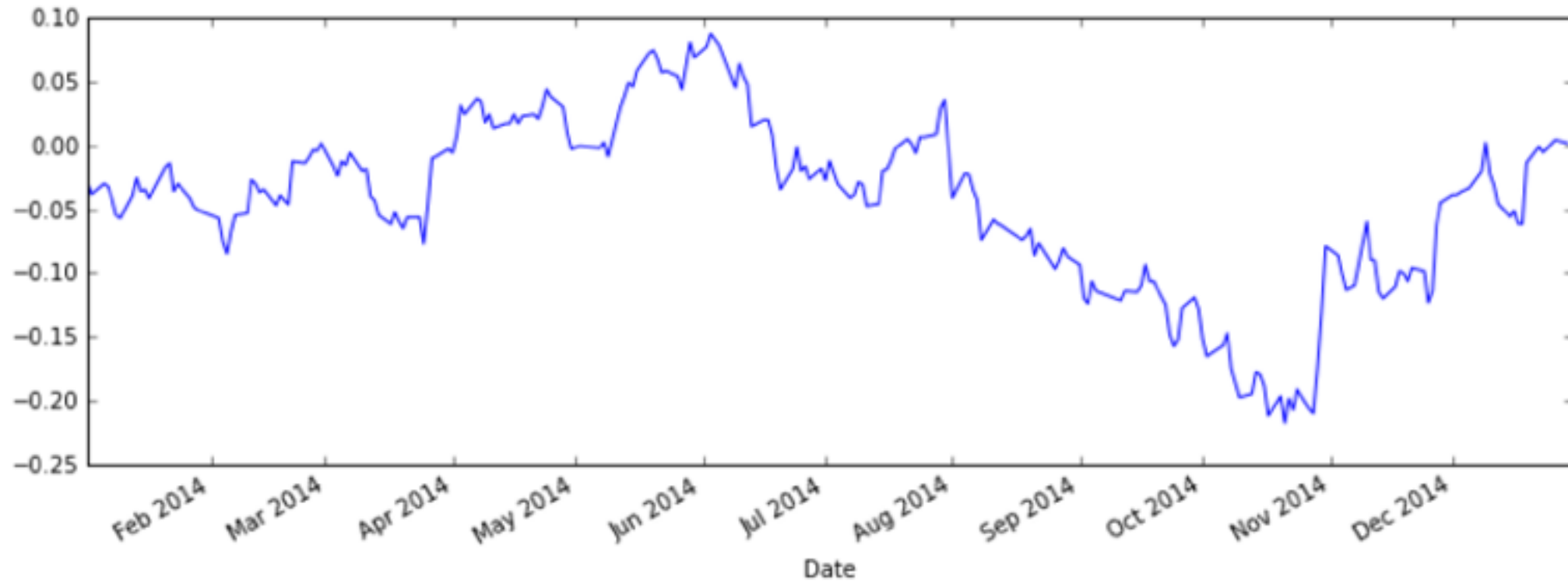

로그 수익률 차트

```
df_005930['Ret'].plot(figsize=(12,4))
```



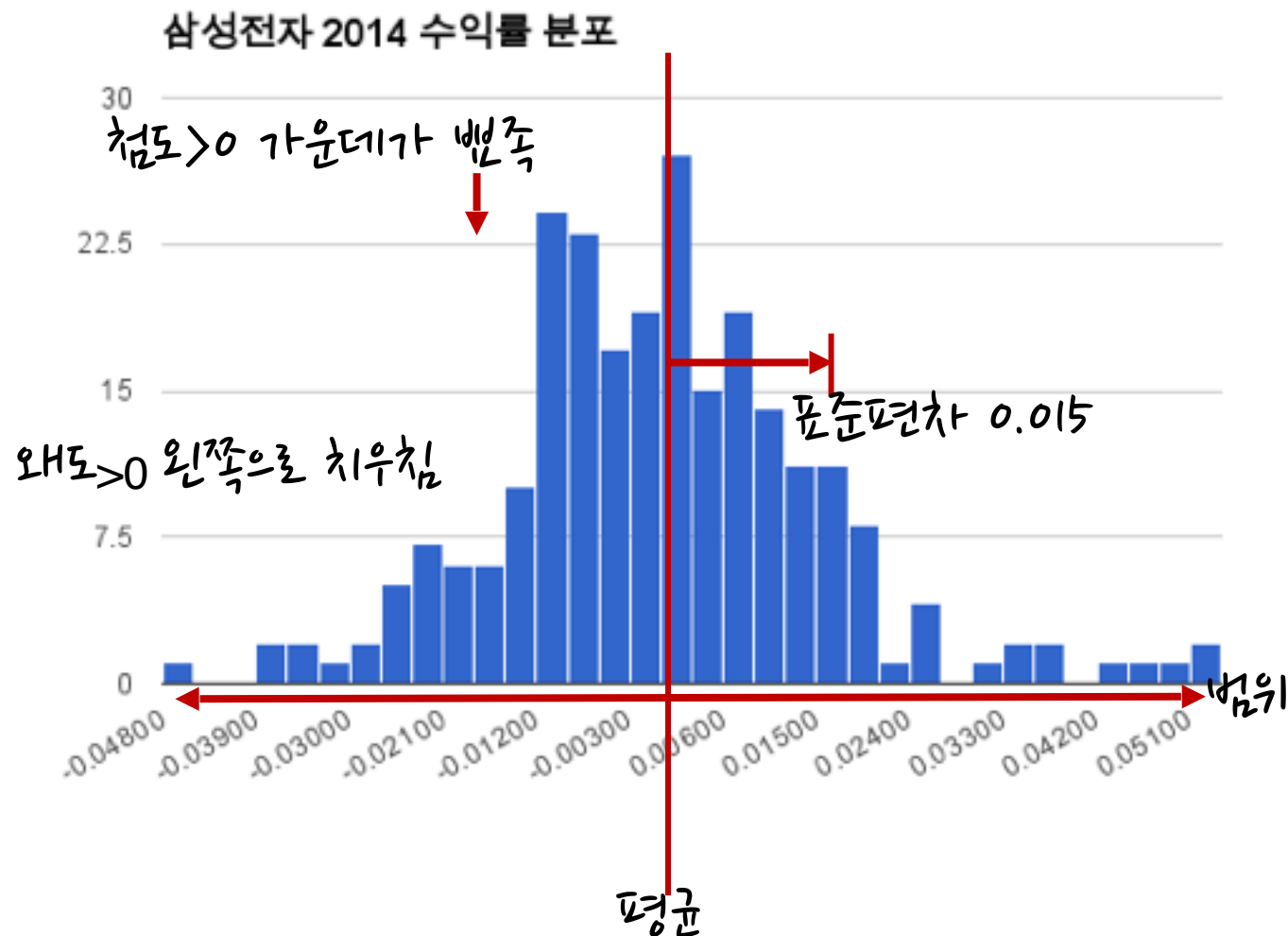
로그 수익률 누적 차트

```
df_005930['Ret'].cumsum().plot(figsize=(12,4))
```



기술통계

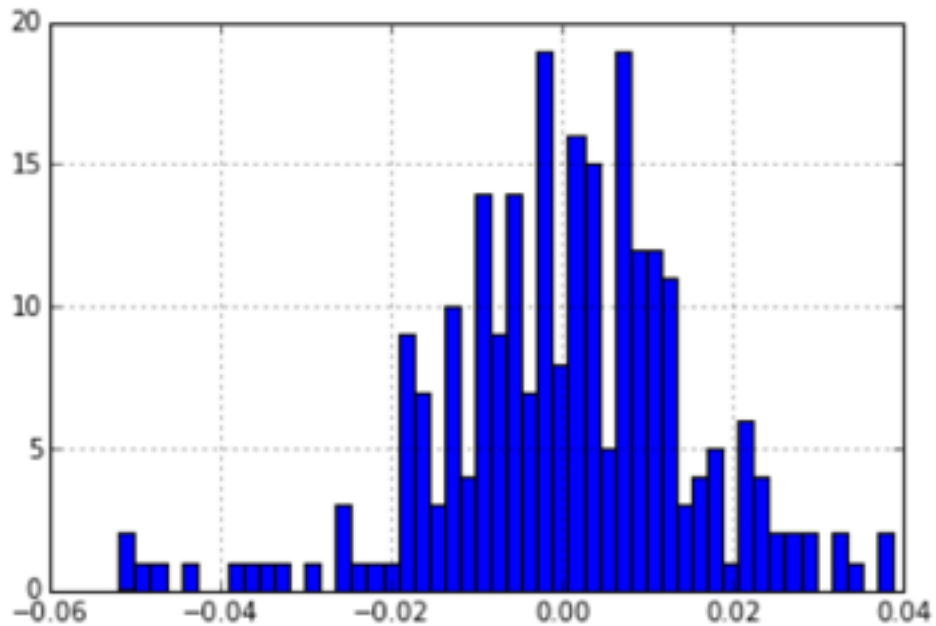
```
def full_describe(df):  
    print ('mean    %15.6f' % (df.mean()))  
    print ('std_err %15.6f' % (df.std() / np.sqrt(df.count())))  
    print ('median  %15.6f' % (df.median()))  
    print ('std     %15.6f' % (df.std()))  
    print ('var     %15.6f' % (df.var()))  
    print ('kurt    %15.6f' % (df.kurt()))  
    print ('skew    %15.6f' % (df.skew()))  
    print ('range   %15.6f' % (df.max() - df.min()))  
    print ('min     %15.6f' % (df.min()))  
    print ('max     %15.6f' % (df.max()))  
    print ('sum     %15.6f' % (df.sum()))  
    print ('count   %15.6f' % (df.count()))
```



평균	-0.000071
표준오차	0.000935
최빈값	0.000000
표준편차	0.015051
분산	0.000227
첨도	1.718902
왜도	0.447813
범위	0.098976
최소값	-0.047006
최대값	0.051970
합	-0.018443
관측수	259

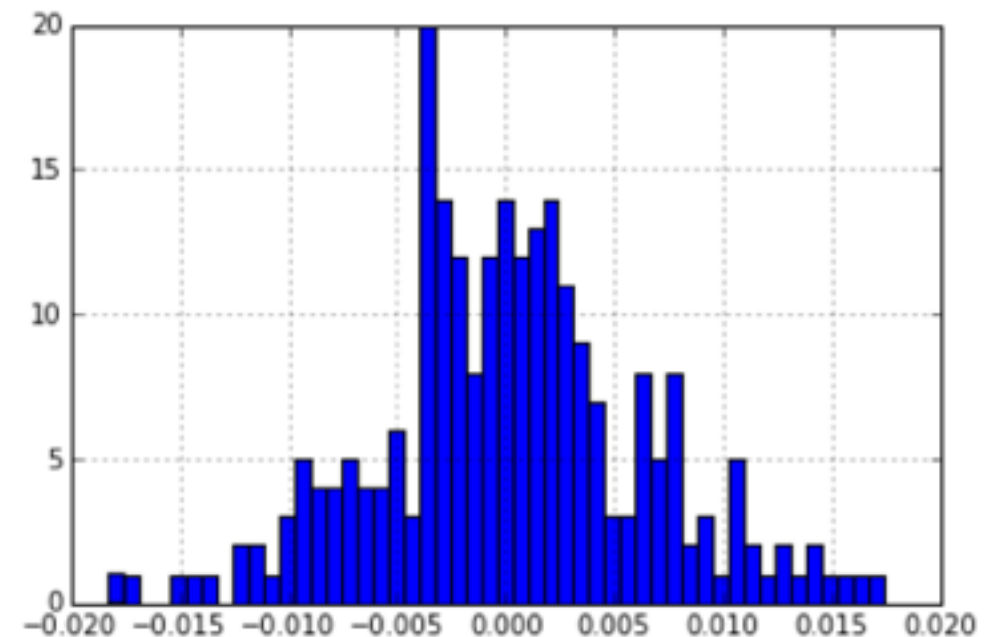
005930.KS

mean	-0.000117
std_err	0.000974
median	0.000786
std	0.015212
var	0.000231
kurt	1.395795
skew	-0.545131
range	0.090685
min	-0.051970
max	0.038715
sum	-0.028563
count	244.000000



^KS11

mean	0.000109
std_err	0.000400
median	0.000013
std	0.006252
var	0.000039
kurt	0.333151
skew	0.081799
range	0.035658
min	-0.018262
max	0.017396
sum	0.026580
count	244.000000



항상 비교되는 지표가 필요하다.

이를 벤치마크라고 하며, 코스피 지수, 업종 지수, 국공채 채권 수익률 같은 대표적인 지표를 사용한다.

왜도, 첨도

- 왜도 *skewness*, 데이터 분포의 비대칭 정도.
0이면 좌우 대칭, 양수면 오른쪽으로 긴꼬리, 음수면 왼쪽으로 긴꼬리.
- 첨도 *kurtosis*, 분포의 정도가 얼마나 중심에 집중해있는지, 분포의 중심의 뽕족한 정도
정규분포에 비해, 0이면 동일, >0 더 뽕족, <0 납작

왜도가 오른쪽 치우치고(-), 첨도가 작을수록(<0) 활발해 지고 변동성이 커짐

분산과 표준편차

- 관측값과 평균의 차이가 "편차" deviation
- "편차" 제곱의 평균이 "분산" variance
- "분산"의 제곱근이 "표준편차" standard deviation

분산^{variance}은 평균을 기준으로 자료들이 얼마나 퍼져 있는지를 재는 척도. (분산=편차 제곱값의 평균)
왜 제곱을 쓰는가? 절대값을 얻기 위해.

“표준편차”가 반이다. “표준편차”만 제대로 알아도 통계의 절반은 먹고 들어간다.

표준편차 standard deviation

- 분산 variance 에 제곱근
- '표준편차'를 대충 '평균적인 편차'라고 생각해도 크게 무리는 없다.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1} (x_i - \mu)^2}$$

- 엑셀에서는 STDEV.P(), STDEV.S() 함수 (엑셀 2010이전: STDEVP(), STDEV())

표준편차의 중요성

금융 통계에서 가장 중요한 것 딱 하나만 꼽으라면
그것은 표준편차

- 표준편차는 얼마나 울퉁불퉁한가를 표현.
- 같은 점수라도 점수가 속한 그룹의 SD의 크기가 중요. 속한 그룹의 SD클 수록 잘한 것이다.
(편차/표준편차 값으로 과목 경쟁력을 비교할 수 있다)
- 평균은 서열, SD는 성질을 표현 (모든 데이터에 a를 더하면 ?)

수익률의 표준편차

투자에서 수익률 뿐만 아니라 표준편차도 중요

- “수익률이 SD의 1배 정도가 될 것이다”라는 생각이 가능
- 수익률의 표준편차 = 주가 변동성 = 거래 리스크(위험지표)
- 위험지표는 역으로 기회지표

샤프지수

Sharpe Ratio, 변동성(위험) 대비 펀드 수익률

- 샤프 지수 = (수익률 - 국공채 수익률) / 수익률의 표준편차
- 초과 수익 대비 수익률의 표준편차
샤프지수가 2이면 위험(SD) 1에 대해 은행이자 대비 2배의 수익이라는 의미
- 높을 수록 좋다(2 이상이면 쓸만한 펀드). 즉, 같은 위험대비 수익을 안정적

마할라노비스 거리 Mahalanobis distance

평균과의 거리가 표준편차의 몇 배인지를 나타내는 값

- 평균이 50, 표준편차가 3인 경우. 56 이란 값의 평균과 거리는 6이다.
마할라노비스 거리는 $(56 - 50) / 3 = 2$. (즉, 표준편차의 2배 거리)
- 어떤 값이 얼마나 일어나기 힘든 값인지, 또는 얼마나 이상한 값인지를 수치화 하는 방법
- 어떤 데이터가 가짜 데이터인지, 아니면 진짜 데이터인지를 구분
- 아주 특이한 데이터가 들어왔다면, 이 값을 제외하고 평균을 내는 것이 더 합리적일 수 있다

요약

- 데이터의 종류: 질적 qualitative data(=categorical), 양적 quantitative data(=numerical)
- 척도 scale: 명목 < 순서 < 구간 < 비율
- 중심경향치 measure of central tendency(평균): 평균값, 중앙값, 최빈값
- 산포도 measure of dispersion: 분산과 표준편차
- 상대적 위치의 측도: 백분위수, 백분율 순위
- 평균값: 산술, 기하, 조화, 가중
- 일반 수익률 & 로그 수익률
- 기본 통계량: 평균, 분산, 표준편차, 왜도/첨도