Statistical Analysis in Python

3장 : 파이썬을 이용한 데이터 분석

3. 파이썬을 이용한 데이터 분석

기술통계

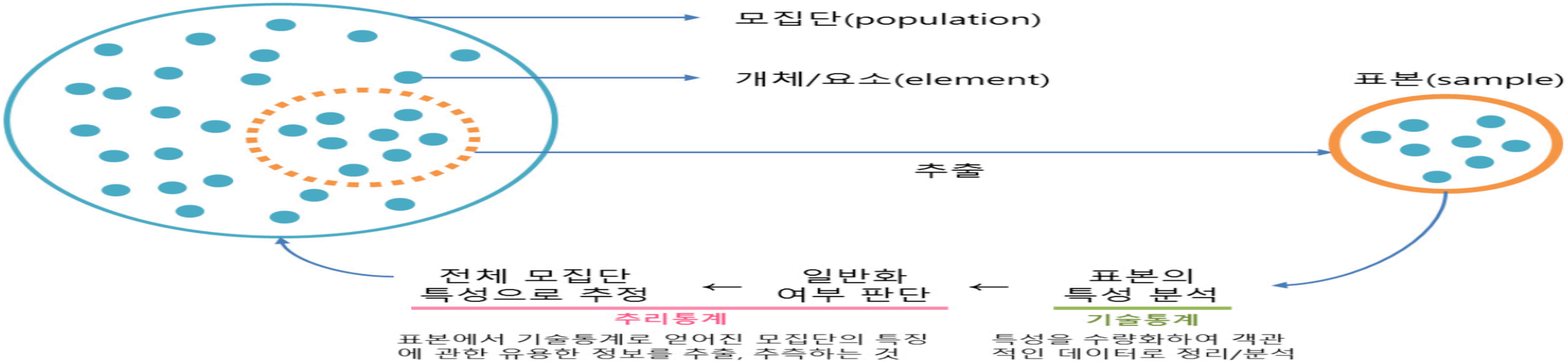
수집한 데이터를 요약, 묘사 설명하는 통계 기법

(=> 대푯값 확인, 데이터 분포 확인 등)

변수에 따른 데이터 구분

1변량 데이터 : 1가지 종류의 데이터 밖에 없을 경우

다변량 데이터 : 여러개의 변수를 조합한 데이터



3. 파이썬을 이용한 데이터 분석

1변량 데이터의 관리

표본평균 $\overline{X}$, 표본분산 S^2, 표본표준편차 S

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\sum_{k=1}^{n} X_k$$

$$S^2 = \frac{1}{n-1}[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2]$$

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

(v) 표본평균을 사용하여 표본분산을 계산한 값임으로 '과소추청'한 경향이 존재한다.
따라서 자유도 (N-1)를 고려하여 추정한것을 불편추정이라고 한다.

3. 파이썬을 이용한 데이터 분석

다변량 데이터 처리 : Tidy Data (깔끔한 데이터)

1) Tidy Data

행정구역별(읍면동)	2017							2018					
	총인구 (명)	남자 (명)	여자 (명)	내국인-계 (명)	내국인-남자 (명)	내국인-여자 (명)	외국인-계 (명)	총인구 (명)	남자 (명)	여자 (명)	내국인-계 (명)	내국인-남자 (명)	내국인-여자
전국	51,422,507	########	########	49,943,260	24,922,392	25,020,868	1,479,247	51,629,512	########	########	49,977,951	24,931,554	25,04
읍부	4,794,377	2,456,558	2,337,819	4,625,994	2,347,613	2,278,381	168,383	4,984,023	2,557,807	2,426,216	4,787,535	2,431,315	2,35
면부	4,835,090	2,498,157	2,336,933	4,592,308	2,324,414	2,267,894	242,782	4,730,290	2,453,691	2,276,599	4,464,067	2,263,139	2,20
동부	41,793,040	########	########	40,724,958	20,250,365	20,474,593	1,068,082	41,915,199	########	########	40,726,349	20,237,100	20,48
서울특별시	9,741,871	4,757,642	4,984,229	9,397,944	4,592,393	4,805,551	343,927	9,673,936	4,718,972	4,954,964	9,299,511	4,536,552	4,76
종로구	157,277	76,670	80,607	146,298	71,757	74,541	10,979	157,967	77,147	80,820	145,911	71,611	7
중구	127,896	62,195	65,701	117,631	57,391	60,240	10,265	129,797	62,975	66,822	117,809	57,240	6
용산구	223,898	108,471	115,427	207,893	99,759	108,134	16,005	226,938	110,685	116,253	207,812	99,806	10
성동구	302,367	149,060	153,307	291,931	144,018	147,913	10,436	306,796	150,623	156,173	295,489	145,220	15
광진구	363,934	177,156	186,778	345,535	168,833	176,702	18,399	362,304	175,762	186,542	342,575	166,730	17
동대문구	357,380	176,460	180,920	339,438	168,977	170,461	17,942	358,141	176,323	181,818	338,293	167,999	17
중랑구	396,892	196,346	200,546	390,879	193,871	197,008	6,013	391,668	193,599	198,069	385,082	190,871	19
성북구	445,417	215,068	230,349	432,226	209,374	222,852	13,191	438,734	211,204	227,530	424,211	205,031	21
강북구	313,698	152,464	161,234	309,255	150,745	158,510	4,443	309,138	150,028	159,110	304,191	148,057	15
도봉구	332,586	161,944	170,642	329,957	160,885	169,072	2,629	328,243	159,569	168,674	325,292	158,370	16
노원구	543,499	263,655	279,844	538,768	261,563	277,205	4,731	534,096	258,737	275,359	528,268	256,124	27
은평구	466,243	225,813	240,430	460,651	223,388	237,263	5,592	462,552	223,452	239,100	456,582	220,818	23
서대문구	321,345	151,132	170,213	308,193	146,320	161,873	13,152	318,874	149,438	169,436	304,508	144,208	16
마포구	368,841	175,326	193,515	356,881	170,311	186,570	11,960	368,181	174,194	193,987	355,405	168,819	18
양천구	452,111	222,503	229,608	446,643	220,035	226,608	5,468	445,591	218,975	226,616	439,910	216,381	22

(Tidy Data)

1) Tidy Data

year 변수

행정구역별(읍면동)	2017							2018					
	총인구 (명)	남자 (명)	여자 (명)	내국인-계 (명)	내국인-남자 (명)	내국인-여자 (명)	외국인-계 (명)	총인구 (명)	남자 (명)	여자 (명)	내국인-계 (명)	내국인-남자 (명)	내국인-여자
전국	51,422,507	########	########	49,943,260	24,922,392	25,020,868	1,479,247	51,629,512	########	########	49,977,951	24,931,554	25,04
읍부	4,794,377	2,456,558	2,337,819	4,625,994	2,347,613	2,278,381	168,383	4,984,023	2,557,807	2,426,216	4,787,535	2,431,315	2,35
면부	4,835,090	2,498,157	2,336,933	4,592,308	2,324,414	2,267,894	242,782	4,730,290	2,453,691	2,276,599	4,464,067	2,263,139	2,20
동부	41,793,040	########	########	40,724,958	20,250,365	20,474,593	1,068,082	41,915,199	########	########	40,726,349	20,237,100	20,48
서울특별시	9,741,871	4,757,642	4,984,229	9,397,944	4,592,393	4,805,551	343,927	9,673,936	4,718,972	4,954,964	9,299,511	4,536,552	4,76
종로구	157,277	76,670	80,607	146,298	71,757	74,541	10,979	157,967	77,147	80,820	145,911	71,611	7
중구	127,896	62,195	65,701	117,631	57,391	60,240	10,265	129,797	62,975	66,822	117,809	57,240	6
용산구	223,898	108,471	115,427	207,893	99,759	108,134	16,005	226,938	110,685	116,253	207,812	99,806	10
성동구	302,367	149,060	153,307	291,931	144,018	147,913	10,436	306,796	150,623	156,173	295,489	145,220	15
광진구	363,934	177,156	186,778	345,535	168,833	176,702	18,399	362,304	175,762	186,542	342,575	166,730	17
동대문구	357,380	176,460	180,920	339,438	168,977	170,461	17,942	358,141	176,323	181,818	338,293	167,999	17
중랑구	396,892	196,346	200,546	390,879	193,871	197,008	6,013	391,668	193,599	198,069	385,082	190,871	19
성북구	445,417	215,068	230,349	432,226	209,374	222,852	13,191	438,734	211,204	227,530	424,211	205,031	21
강북구	313,698	152,464	161,234	309,255	150,745	158,510	4,443	309,138	150,028	159,110	304,191	148,057	15
도봉구	332,586	161,944	170,642	329,957	160,885	169,072	2,629	328,243	159,569	168,674	325,292	158,370	16
노원구	543,499	263,655	279,844	538,768	261,563	277,205	4,731	534,096	258,737	275,359	528,268	256,124	27
은평구	466,243	225,813	240,430	460,651	223,388	237,263	5,592	462,552	223,452	239,100	456,582	220,818	23
서대문구	321,345	151,132	170,213	308,193	146,320	161,873	13,152	318,874	149,438	169,436	304,508	144,208	16
마포구	368,841	175,326	193,515	356,881	170,311	186,570	11,960	368,181	174,194	193,987	355,405	168,819	18
양천구	452,111	222,503	229,608	446,643	220,035	226,608	5,468	445,591	218,975	226,616	439,910	216,381	22

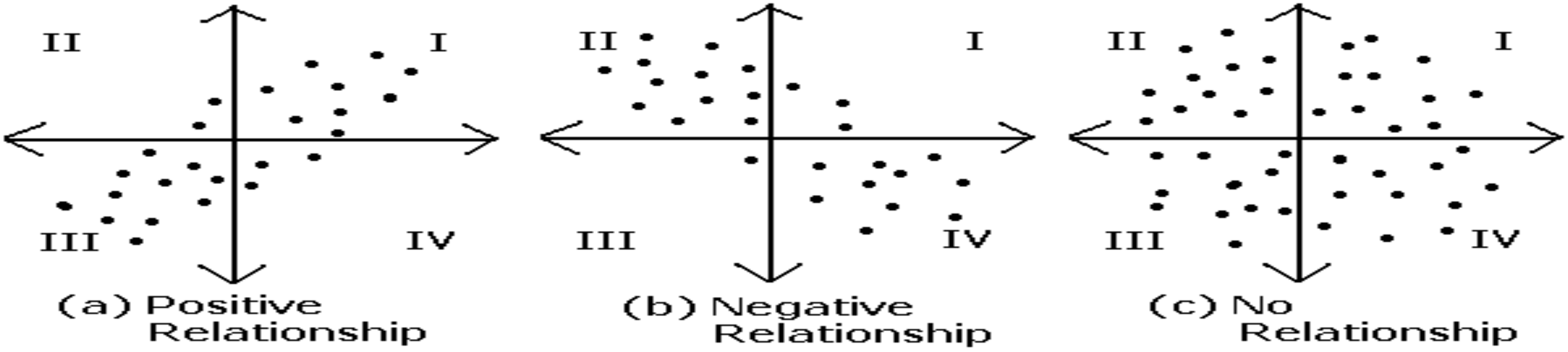
(Tidy Data)

3. 파이썬을 이용한 데이터 분석

다변량 데이터 관리

공분산 : 2개의 연속형 변수의 관계성을 확인하는 통계량
공분산의 결과가 0보다 큰 값일 때는
'변수 x가 평균보다 큰 값일 때 다른 변수 y도 평균보다 큰 값을 갖는다고 기대할수 있다'

$$Cov(x,y) = \frac{\sum (x-\overline{x})(y-\overline{y})}{n}$$

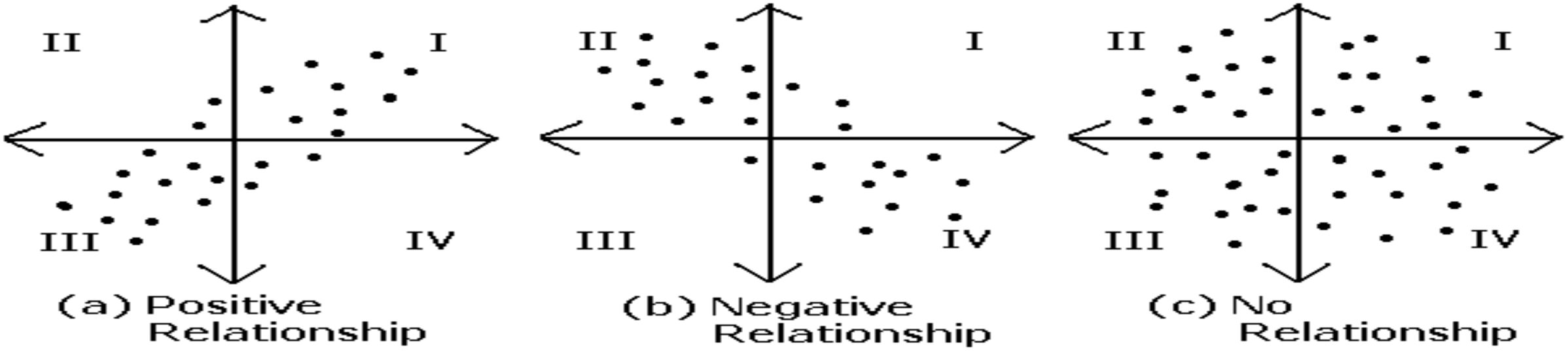


3. 파이썬을 이용한 데이터 분석

다변량 데이터 관리

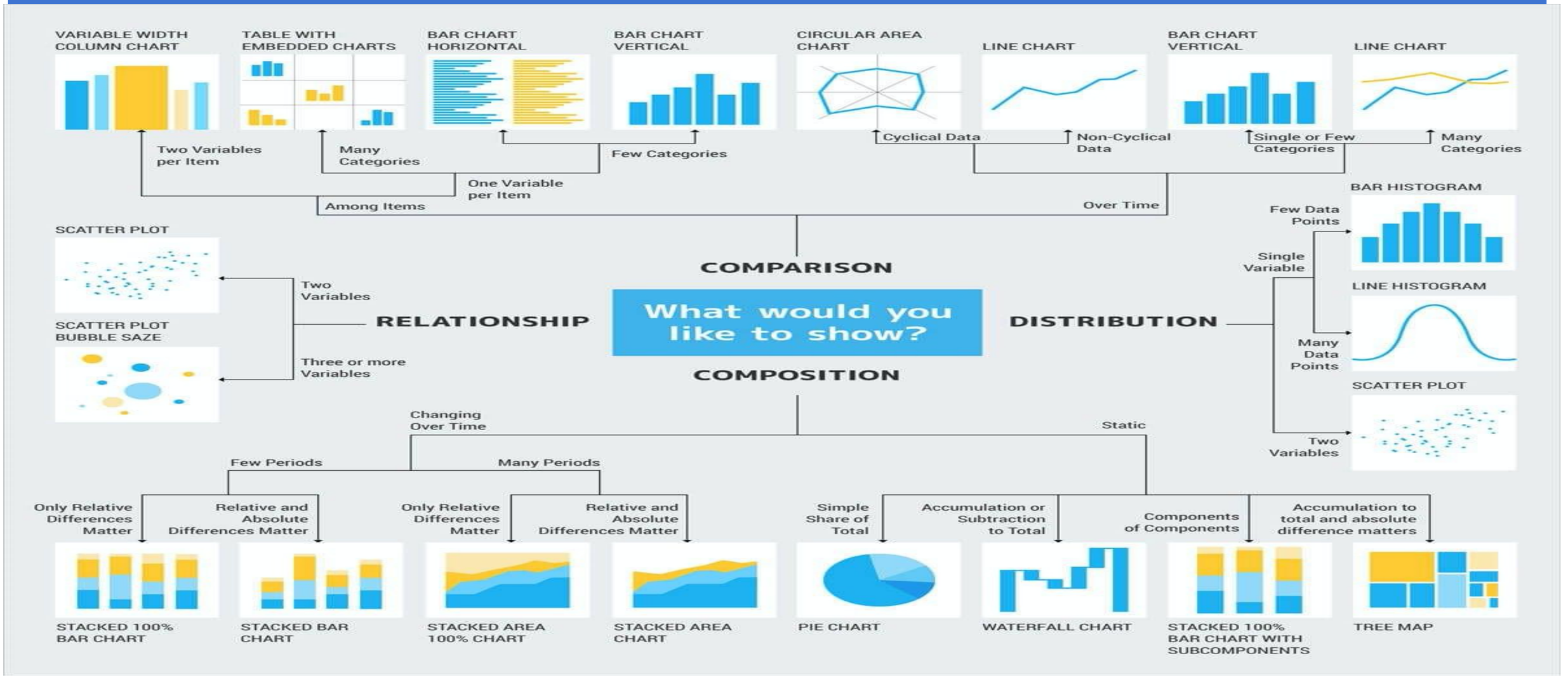
피어슨 상관계수 : 공분산을 최대값 1, 최솟값 -1 사이가 되도록 표준화하는 것

$$r_{XY} = \frac{\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i^n \left(X_i - \overline{X}\right)^2}\sqrt{\sum_i^n \left(Y_i - \overline{Y}\right)^2}}$$



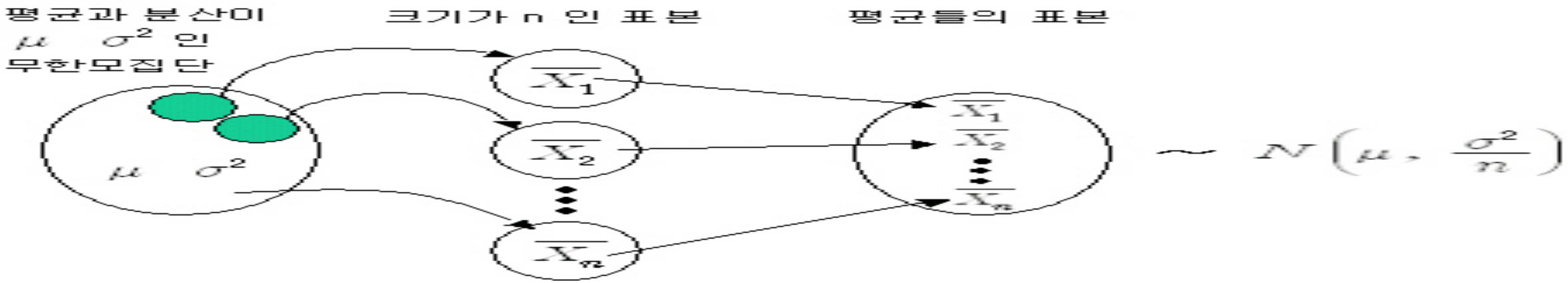
3. 파이썬을 이용한 데이터 분석

다변량 데이터 관리 & 분포 파악



3. 파이썬을 이용한 데이터 분석

모집단분포를 정규분포로 가정해도 좋은가 : "중심극한 정리"



모집단분포를 정규분포로 가정해도 좋은가

샘플 사이즈가 크면 표본평균은 모평균에 가까워진다. (일치성)

(일치성 : 샘플사이즈가 커지면 추정량이 진짜 모수에 가까워지는 특징)

표준오차(Standard Error : SE) : 표준편차 / squirt(표본수)

3. 파이썬을 이용한 데이터 분석

모집단분포를 정규분포로 가정해도 좋은가

모집단분포를 알면 모르는 데이터에 대한 예측 및 추측을 할수 있게 된다.

모집단분포를 정규분포로 가정하면 정규분포의 모수를 추정하는 것으로 모집단분포를 추정 가능