상관분석(Correlation Analysis)

http://blog.naver.com/PostView.nhn?blogId=y4769&logNo=220227007641

상관분석은 회귀분석을 하기 위한 전초전으로서 비교적 쉬운 분석 중의 하나이다. 어떻게 보면 연구자들이 가장 많이 시행하는 t-test 보다도 더 쉬운 분석이 아닐까 생각해본다. 물론 분석 방법 뿐만아니라 결과 해석까지 어렵지 않으니 한번만이라도 제대로 따라해보길 바란다.

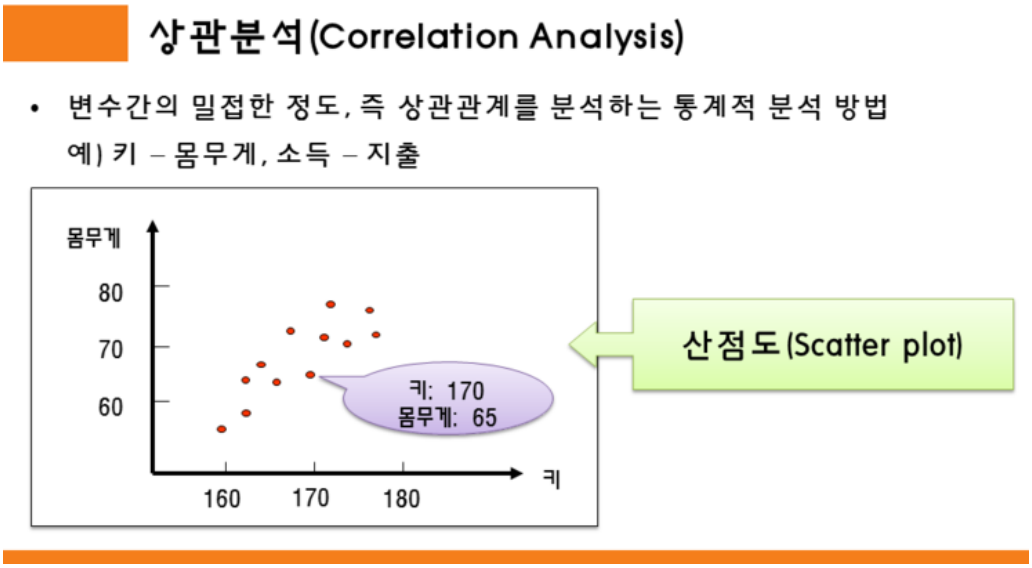
정의

- 연속 변수로 측정된 두 변수간의 선형 관계를 분석하는 기법
- 한 변수가 증가하면 다른 한 변수도 선형적으로 증가 혹은 감소하는지를 나타낸 것
- 상관계수(Correlation Coefficient ; ρ [rho])
  - ➔ 두 변수 사이의 선형적인 관계 정도를 나타냄
  - ➔ 통계학에서 Pearson 상관계수를 의미

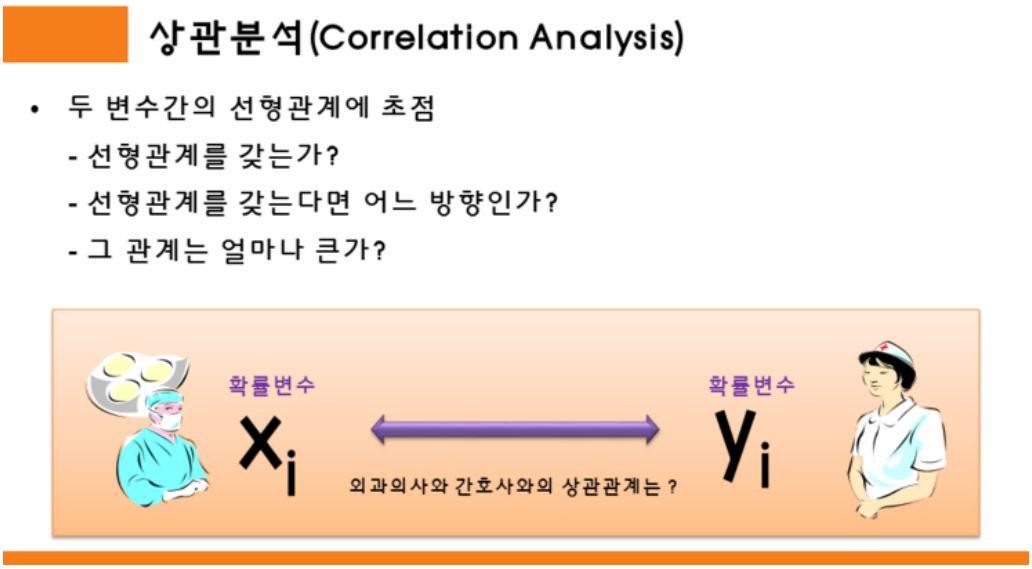
상관분석은 연속형 변수로 측정된 두 변수 간의 선형적 관계를 분석하는 기법이다. 연속형 변수는 산술 평균을 계산할 수 있는 숫자형의 데이터이며, 선형적 관계라 함은 흔히 비례식이 성립되는 관계를 말한다. 예를들어 A 변수가 증가함에 따라 B 변수도 증가되는지 혹은

감소하는지를 분석하는 것이다.

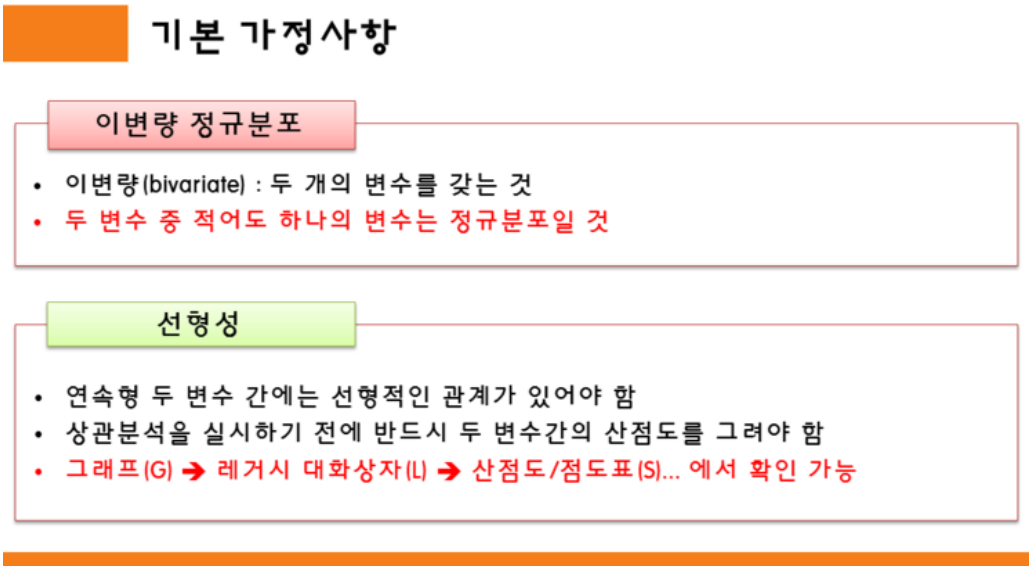
상관분석에는 두 변수 사이의 선형적인 관계 정도를 나타내기 위해 상관계수(correlation coefficient)를 사용한다. 상관분석에는 측정 데이터에 따라 피어슨 상관분석, 스퍼만 상곤분석 등의 여러가지 분석 방법이 있지만, 일반적으로 상관계수라 함은 피어슨 상관계수 (Pearson correlation coefficient)를 의미한다.



위에서 언급했듯이 상관분석 두 변수간의 밀접한 정도, 즉 상관관계를 분석하는 통계적 분석 방법으로, 키와 몸무게를 통해 예를들어 보면, 위와 같은 그래프를 얻을 수 있을 것이다. 위의 그래프를 통해 키가 증가할수록 몸무게 또한 증가하는 것을 볼 수 있으며, X와 Y축에 점으로 표시한 그래프를 산점도라 일컫는다. 상관분석을 하기 위해서는 반드시 산점도를 통하여 선형성을 확인하여야 한다.



다시한번 말하지만, 상관분석은 두 변수간의 선형적 관계에 초점을 두고 있으며 선형 관계를 갖는가? 어느 방향으로 형성되어 있는가? 그 관계의 정도는 얼마나 큰가? 와 같은 질문을 던지며 결과 해석을 해볼 수 있겠다.



모수통계는 일단, 데이터의 기본 가정사항을 확인해야 된다.

상관분석의 가정사항은 이변량 정규분포와 두 변수 사이의 선형성을 충족하면 된다.

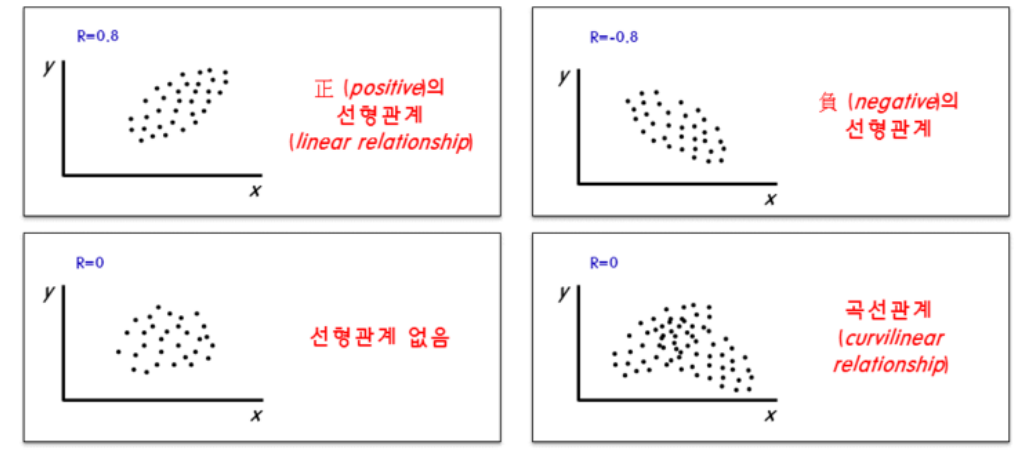
이변량(bivariate)이란 두 개의 변수를 말하는 것이다.

두 변수 모두 정규분포를 하면 가장 좋겠지만, **적어도 하나의 변수만이라도 정규성을 만족하면 정규성에 대한 가정사항은 충족시킨다**고 한다. 그런데 실제 분석에서 보면 정규성을 만족하지 못하는 경우가 많아 혹자는 현실적으로는 정규성 분석을 시행하지 않을 수도 있다고 한다. 어쨌든 제시된 기본가정 사항 중의 하나가 정규분포이므로 정규성 검정을 먼저 시행하고 다음으로 진행하도록 한다.

만약, 두 변수 모두 정규성을 만족하지 못한다면 Pearson 상관분석의 비모수검정에 해당하는 **Spearman Correlation Analysis(스퍼만 상관분석)**을 시행한다.

다음, 정규성을 만족한 경우 선형성을 확인하기 위해 산점도를 확인해야 하는데, 이는 SPSS의 산점도/점도표(S)... 의 메뉴를 통해 쉽게 확인이 가능하다.

두 변수의 관계



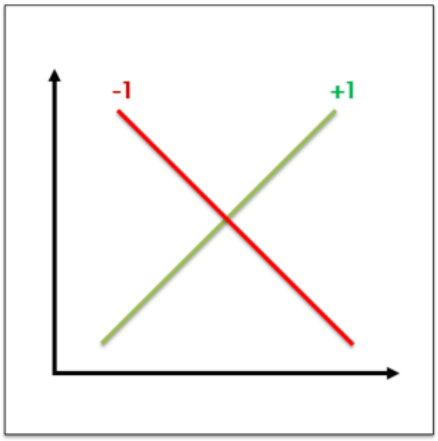
산점도를 통해 나온 그래프를 살펴보자.

먼저 위의 왼쪽 그패프는 X가 증가함에 따라 Y도 같이 증가하는 정의 선형관계 또는 양의 선형관계를 나타내고 있다.

위의 오른쪽 그래프는 반대로서 부의 선형관계 또는 음의 선형관계를 나타내고 있다.

그리고 밑의 두 그래프는 선형적 관계가 없는 그래프로서, 이때는 피어슨 상관분석을 시행할 수 없다.

상관계수(Correlation coefficient)



- 상관관계의 크기를 나타내는 값
- -1 ~ +1 사이의 값
  - ➔ +1 = 가장 높은 양의 상관관계
  - ➔ -1 = 가장 높은 음의 상관관계
  - ➔ 0 = 상관관계 없음

피어슨 상관계수(Pearson Correlation Coefficient)

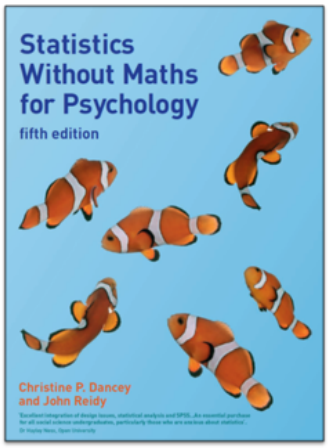
$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 (Y_i - \overline{Y})^2}}$$

상관 관계의 크기를 나타내는 값을 상관계수라고 표현하며, -1 부터 +1 사이의 값을 가진다. 이때 +1 또는 -1의 경우 상관관계가 가장 크며, 0으로 나올 경우 상관관계가 전혀 없음을 나타낸다.

Correlation Coefficient

Perfect	+1	-1
Strong	+0.9 +0.8 +0.7	-0.9 -0.8 -0.7
Moderate	+0.6 +0.5 +0.4	-0.6 -0.5 -0.4
Weak	+0.3 +0.2 +0.1	-0.3 -0.2 -0.1
Zero	0	

Christine Dancey and John Reidy, Statistics Without Maths for Psychology, p.175. Prentice Hall, 5th edition, 2011.

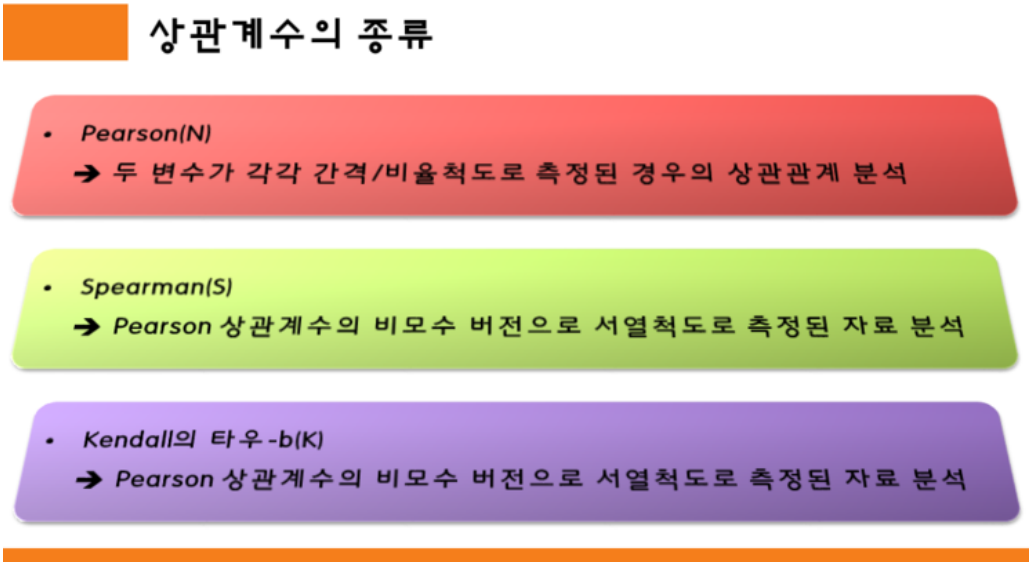


Dancey 등은 상관계수에 따라 Zero부터 Perfect까지 등급을 나뉘었으며, 논문의 고찰에서 결과 해석시 매우 유용하게 쓰일 수 있는 자료이므로 기억해두록 한다.

상관관계의 종류와 자료

종류	척도
Pearson 상관관계	간격/비율 – 간격/비율
Spearman 서열상관관계	서열 – 서열
Kendall의 tau	서열 – 서열
Point biserial r	간격/비율 – 명목(2분변수)
phi coefficient	명목(2분변수) – 명목(2분변수)

그런데 상관분석에는 측정된 데이터에 분석 방법이 달라진다. 지금까지 얘기한 피어슨 상관관계는 간격 척도 및 비율 척도와 같은 연속형의 데이터에 적용되며, 그외 서열척도에는 스퍼만 상관분석이나 켄달의 타우 상관분석이 주로 이용된다.



피어슨 상관분석은 연속형 데이터, 스퍼만 상관분석 및 켄달의 타우는 서열척도로 측정된 데이터 분석에 이용되므로 꼭 기억해둔다.