Lecture 05 - 상관 관계 파악하기

개요

- 데이터의 종류: 연속형, 이산형
- 단변수 요약: 도수분포표, 히스토그램 이해
- 상관 관계 파악: 변수 2개의 선형 관계 이해
- 단순 회귀 분석 이해

자료 요약 예시: 나폴레옹 러시아 침공

https://en.wikipedia.org/wiki/French_invasion_of_Russia#/media/File:Minard.png

- 단순한 자료의 나열은 별 다른 정보를 주지 못하기 때문에 다양한 방법으로 요약할 필요가 있음

데이터의 종류

연속형(continous) 자료

- 두 값 사이에 반드시 새로운 값이 존재할 수 있는 자료형
    - 170cm, 171cm 사이에는 170.5cm가 존재.
    - 170cm, 170.5cm 사이에는 170.25cm가 존재.
- 무한히 확장 가능함.
- uncountable

이산형(discrete) 자료

- 두 값 사이에 새로운 값이 존재하지 않는 경우가 있는 자료형
    - 1개와 3개 사이에는 2개가 존재.
    - 1개와 2개 사이에는 1.5개 없음.
- finite or countably infinte
- 범주형(categoical) 자료

명목형(nomial)

- 이산형 자료 중, 순서 구분이 없는 자료형
    - 성별, 나라 등

순서형(ordinal)

- 이산형 자료 중, 순서 구분이 있는 자료형
    - 성적(A>B>C>D>F), 5점척도(5>4>3>2>1)

```
R: factor
factor(x = character(), levels, labels = levels,
      exclude = NA, ordered = is.ordered(x), nmax = NA)
```

자료형 변환

연속형 자료를 이산형으로 바꾸는 경우

- 나이(연속형) -> 나이대(10대, 20대, 30대…)

이산형 자료를 연속형으로 바꾸는 경우

- 돈(이산형) -> 돈(연속형)

도수분포표(frequency table)

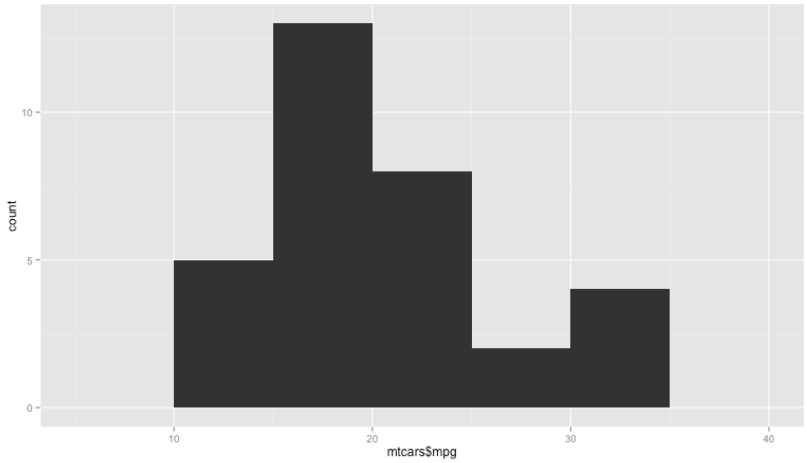
- 이산형 자료 중에 어떤 값이 가장 많은지 세어서 정리한 표

```
> mtcars$cyl
 [1] 6 6 4 6 8 6 8 4 4 6 6 8 8 8 8 8 8 4 4 4 4 8 8 8 8 4 4 4 8 6 8 4
> table(mtcars$cyl)
 4  6  8
11  7 14
```

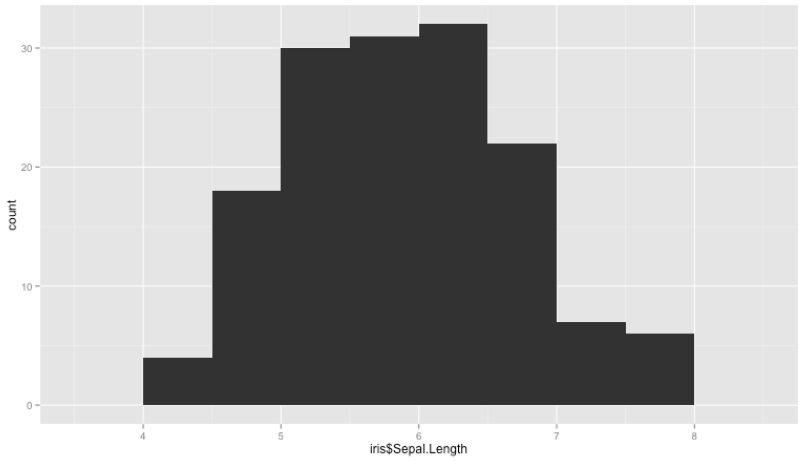
히스토그램(histogram)

- 도수 분포표를 막대 그래프로 표현한 그림

```
> library(ggplot2)
> qplot(mtcars$mpg, binwidth=5)
```




```
> qplot(iris$Sepal.Length, binwidth=.5)
```



상자 그림(box plot):

http://en.wikipedia.org/wiki/Box_plot

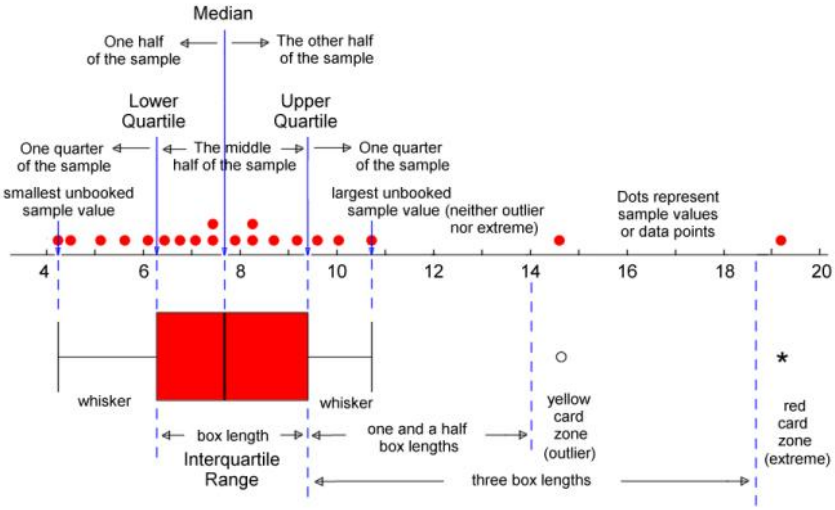


그림 출처: http://web.pdx.edu/~stipakb/download/PA551/boxplot.html

Box Plot vs. histogram

- 히스토그램이 더 많은 정보를 보여주지만, 더 복잡하다.
- 박스 플롯이 더 간결하지만, 자료 분포를 많이 요약한다.

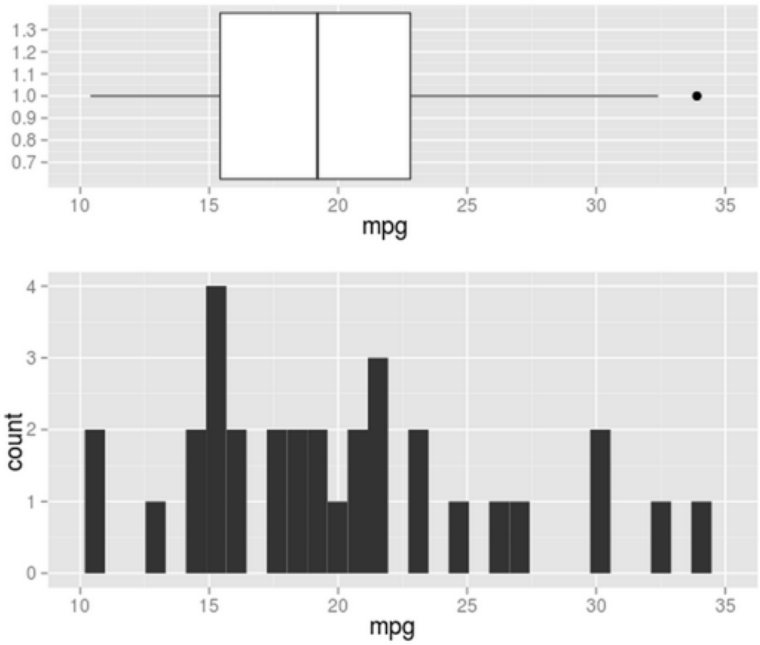


그림 출처:
http://stackoverflow.com/questions/4551582/combination-boxplot-and-histogram-using-ggplot2

상관관계 파악

- 두 변수 사이의 관계
- 산점도(scatter plot)로 먼저 파악
- 상관 계수 구하기

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

산점도 그리기

- mammals data:
  - brain(g)
  - body(kg)
  - 필요하다면 변수 변환(log)

상관 계수(correlation coefficient)

- 상관 계수에는 여러 종류가 있음.
- 대표적으로 '피어슨 상관 계수(Pearson's correlation coefficient')를 사용.
- 범위: -1 과 1 사이
- 관계가 클수록 절대값이 커진다.

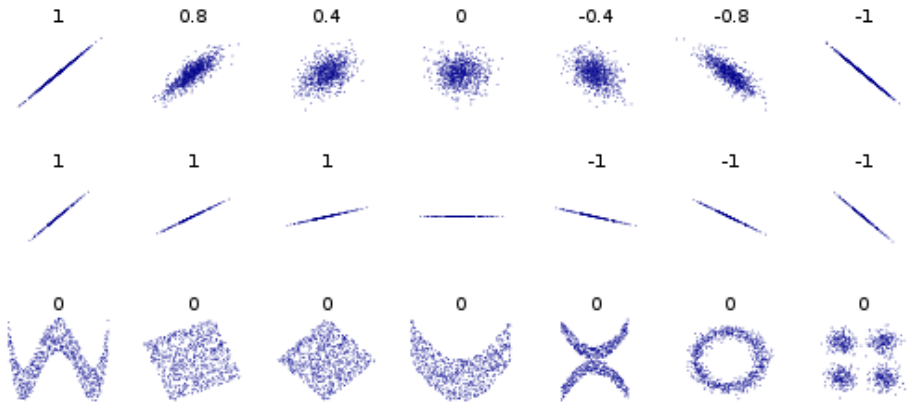


그림 출처: http://en.wikipedia.org/wiki/Correlation_and_dependence

상관 관계와 인과 관계

- 상관 계수가 높다고 인과 관계가 성립하지 않는다.
- 예시:
  혈압과 연봉의 관계
  아이스크림 판매량과 상어에게 공격 받은 피서객 수
  인구10만명당 경찰 수와 범죄율
  맥주 소비량과 영아 사망률
- 인과 관계 파악을 위해서는 실험이 필요!

상관 관계와 선형 관계

- 상관 계수는 선형성(linearity)에 대한 지표

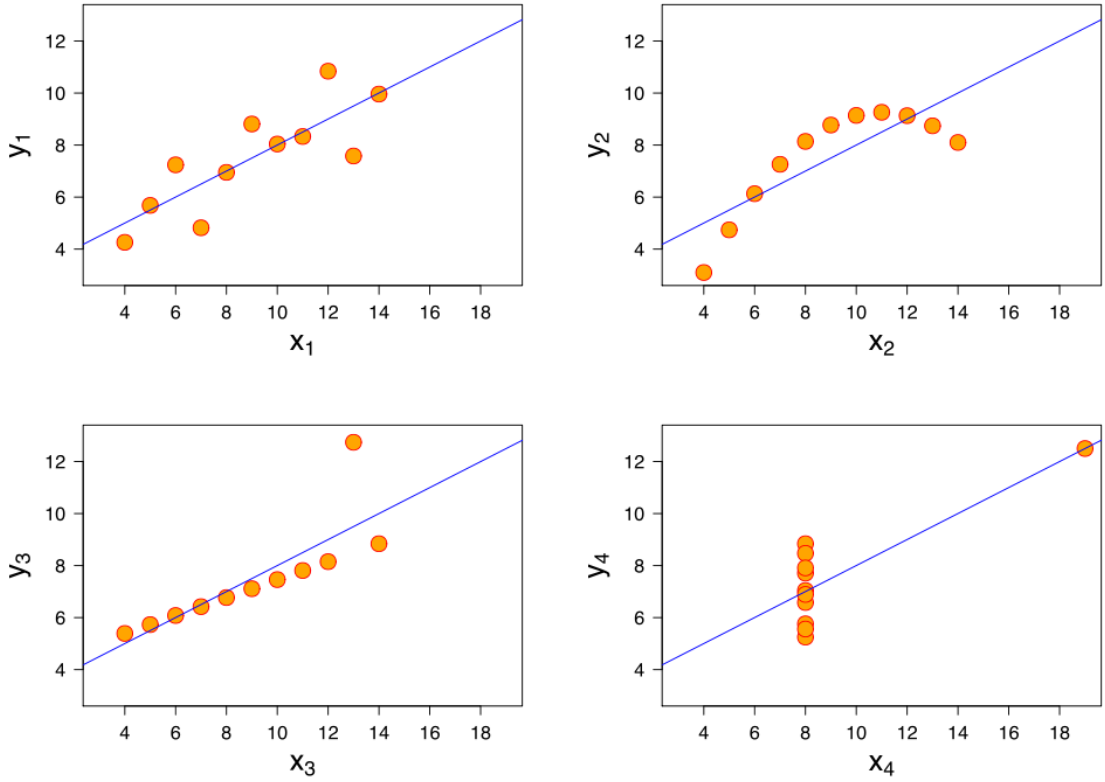
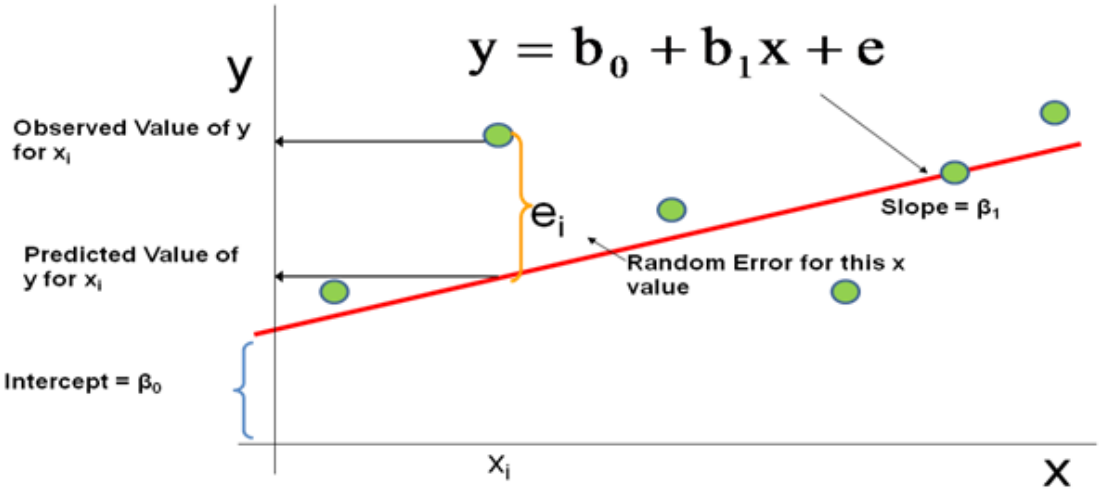


그림 출처:
http://en.wikipedia.org/wiki/Correlation_and_dependence#Correlation_and_linearity

단순 회귀 분석(simple regression analysis)

- y(종속 변수)를 x(설명 변수)로 설명하려는 시도
- y = b0 + b1 * x + e
- sum(e^2)를 최소화하는 직선을 찾는다.
    - 최소 자승법 (least-squares approach)
- 얼마나 x가 y를 잘 설명하나?
    - R^2



Geometric view of OLS regression

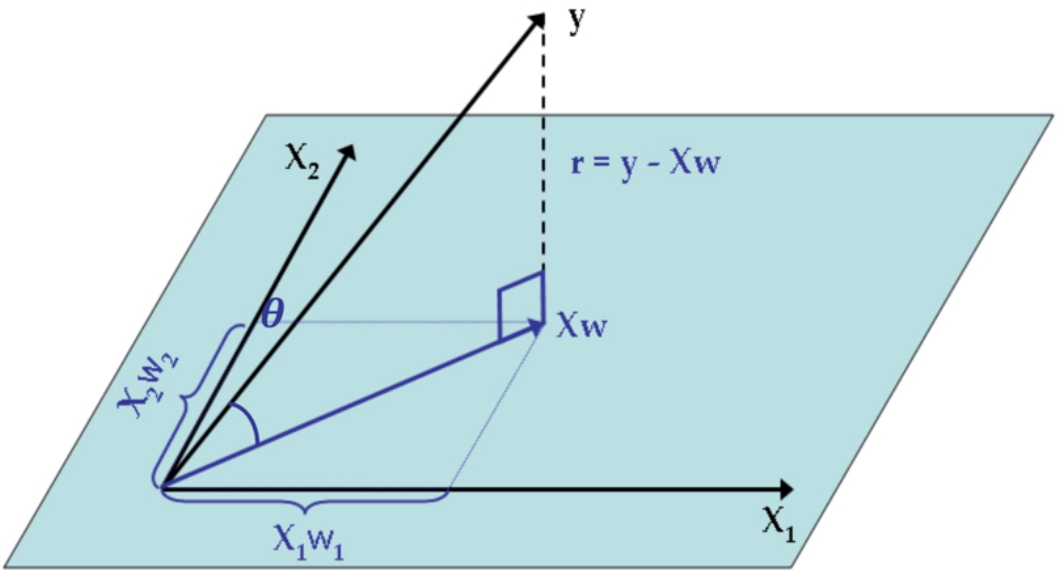


그림 출처: http://learning.cis.upenn.edu/cis520_fall2009/index.php?n=Lectures.Regression