분석프로그래밍 I.07
데이터 시각화

국민대학교 경영학부 빅데이터경영통계학과

2018. 4. 15(월)

1 지난 시간 복습

1.1 데이터 합치기

- 행 합치기
  - `rbind(df1, df2)`
  - `dplyr::bind_rows(df1, df2)`
- 열 합치기
  - `cbind(df1, df2)`
  - `merge(df1, df2, by= )`
  - `dplyr::full_join(df1, df2, by= )`

1.2 Long-form, wide-form 형태 바꾸기

- 패키지 `reshape2`
  - `melt(df, id= ), dcast(df, id ~ measure)`
- 패키지 `tidyr`
  - `gather(df, id= ), spread(df, key= , value= , ... )`

표 1: 여러가지 요약 통계치

집중경향치	최대값, 최소값, 평균, 중앙값 등
변산성	분산, 표준편차, IQR 등
분포의 모양	왜도, 첨도 등
두 변수의 연관	상관계수, 회귀계수 등
이상치의 존재	최대값, 최소값 등

2 데이터를 알아가는 두 가지 방법

- 여러 가지 요약 통계치을 구해 본다(표 2).
- 데이터를 시각화한다(Listing 1). 요약 통계치에서 얻을 수 없는 정보를 얻을 수 있다!

```
data(anscombe)
dat1 <- rbind(anscombe[,c(1,5)])
dat2 <- rbind(anscombe[,c(2,6)])
dat3 <- rbind(anscombe[,c(3,7)])
dat4 <- rbind(anscombe[,c(4,8)])

colnames(dat1) <- c("x","y")
colnames(dat2) <- c("x","y")
colnames(dat3) <- c("x","y")
colnames(dat4) <- c("x","y")

dat1$source <- 1
dat2$source <- 2
dat3$source <- 3
dat4$source <- 4

dat <- rbind(dat1, dat2, dat3, dat4)

by(dat, dat$source, summary)
by(dat, dat$source, psych::describe)
by(dat[,c(1,2)], dat$source, psych::describe)
by(dat[,c(1,2)], dat$source, psych::corr.test)
by(dat[,c(1,2)], dat$source, plot)

```

```
25
26 library(ggplot2)
27 dat$source = factor(dat$source)
28 ggplot(dat, aes(x=x,y=y,col=source)) + geom_point()
29 ggplot(dat, aes(x=x,y=y)) + geom_point() + facet_wrap(~source)
30 ggplot(dat, aes(x=x,y=y, col=source)) + geom_point(size=2) +
31   facet_wrap(~source)
32 ggplot(dat, aes(x=x,y=y, col=source)) + geom_point(size=2) +
33   geom_smooth(method="lm") +
34   facet_wrap(~source)
35 ggplot(dat, aes(x=x,y=y, col=source)) + geom_point(size=2) +
36   geom_smooth(method="lm") + theme_bw() +
37   facet_wrap(~source)
```

Listing 1: `anscombe` 데이터 시각화

3 시각화의 원리 (Grammar of Graphics)

- `geom` : 무엇으로 나타낼 것인가?
- `aesthetics` : 어떤 시각적 특성 (visual attributes) 을 사용할 것인가?

  - 위치 (`x, y`), 크기 (`size`), 모양 (`shape`), 색깔 (`col`), 농도 (`alpha`) 등
- 측정 척도의 종류에 맞는 시각적 특성을 선택한다 (표 2). 자료에 담기지 않은 정보를 시각화하지 않도록 주의한다.[1]

표 2: 측정 척도의 종류와 시각적 특성

측정 척도	특징	가능한 연산	예	변환 가능한 시각적 특성
명목Nominal		`=, !=`	사람이름, 나라이름	색깔, 모양
서열Ordinal		`..., >, <`		
간격Interval	임의적인 0	`..., +, -`	날짜, 위치, 온도	위치, 크기, 농도
비율Ratio		`..., *, -`	체중, 절대온도	위치, 크기, 농도

[1] 참조 : Line Chart on a Non-Continuous Axis (`http://livingqlikview.com/the-9-worst-data-visualizations-ever-created/`)

- 중요한 정보는 정확하게 인식할 수 있는 시각적 특성을 활용한다.[2]

  - 시각적 특성 인식의 정확성 : 위치 > 길이> 각도> 면적> 부피> 색깔> 밝기.[3]

4 패키지 ggplot2

https://www.rstudio.com/wp-content/uploads/2015/05/ggplot2-cheatsheet.pdf

[2]파이 차트(pie chart)를 사용하지 마라!
https://www.geckoboard.com/blog/pie-charts/#.WtDawYhuaUk
https://www.quora.com/How-and-why-are-pie-charts-considered-evil-by-data-visualization-experts

[3]참고 : http://hci.stanford.edu/courses/cs448b/f09/lectures/CS448B-20091005-Perception.pdf