KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

10. 연구방법의 적용 III. 중심경향성 및 관계 파악

정치학연구방법론

박상훈 (sh.park.poli@gmail.com)
강원대학교

연구방법의 적용 III. 중심경향성 및 관계 파악

측정 수준(levels of measurement)

- 명목형(Nominal)
  - 변수에 속한 값들이 서로 배타적이며(mutually exclusive), 구별이 가능하며, 순서를 가지지 않음.
  - 이항(binary) 변수 또는 이산(dichotomous) 변수
- 순서형(Ordinal)
  - 변수에 속한 값들이 순위/순서를 가짐.
  - 다만 상대적 순위만을 알려줄 뿐, 순위 간의 차이가 정확하게 어느 정도인지는 알려주지 않음.

연구방법의 적용 III. 중심경향성 및 관계 파악

- 등간형(Interval)
  - 변수의 값들 간의 차이가 서로 일정한 간격을 가짐.
- 비율형(Ratio)
  - 절대영(absolute zero)이 존재
  - 예) 경제성장율 0%일 경우, 경제가 '전혀' 성장하지 않았다는 것을 의미. 반면 등간형 변수인 온도가 0도라고 해서 온도가 존재하지 않는 것은 아님.

명목형-, 순위형-, 그리고 등간형 데이터는 보통 이산형일 가능성이 크고, 비율형 데이터는 연속형 자료일 가능성이 큼.

- 다만 학자들은 대개 등간형과 비율형 데이터를 거의 같은 것처럼 취급하고는 함.

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

평균(Mean)

$$\bar{x} = \frac{\sum_{i=1}^{n}}{n}$$

- 일종의 균형점이라고 생각하면 이해가 편함.
- 값들 간의 거리 차이의 제곱을 최소화한 결과라고 생각할 수 있음.
- 위와 같은 평균을 산술평균(arithmetic mean)이라 하며 이외에도 기하평균(geometric mean), 조화평균(harmonic mean) 등이 있음.

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

평균(Mean)

- 기하평균: $\bar{x} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$
- 조화평균: $\bar{x} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}}$

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

평균(Mean)

그렇다면 평균을 왜 보는 것일까?

- 평균은 중심(central)을 보여줄 수 있는 하나의 지표에 지나지 않음.
- 예를 들어, A, B, C 학급의 영어 실력을 비교하고 싶다고 하자.
    - 아무런 추가 정보가 없을 때, 우리는 각 학급의 영어 실력을 무엇으로 파악할까?
    - 평균이란 어떠한 집단의 정보를 요약하여 그것을 대표하는 값이라는 의미를 가짐.
    - 우리는 잘 모를 때, 그나마 틀릴 가능성이 제일 낮은 값---평균을 제시하곤 함.

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

평균(Mean)

그렇다면 평균을 왜 보는 것일까?

- 만약 90점이 8명, 20점이 2명인 학급이 있다고 하자. 평균은 얼마일까?
    - 76점---이 평균이 과연 이 학급의 영어 실력을 잘 보여준다고 할 수 있을까?

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

중앙값(Median) & 최빈값(Mode)

만약 평균이 우리가 기대하는 대표값으로서 제대로 기능하지 못한다면 어떻게 될까?

- 평균에 대한 대안들: 중앙값과 최빈값
- 중앙값: 말 그대로 중앙에 놓인 값. 5개의 값이 있다면 3번째에 해당하는 값이 중앙값이라고 할 수 있음.
    - 구체적으로는 편차(deviations)의 절대값의 합이 최소가 되게 하는 x의 값
    - 평균과는 다르게 이탈치(outliers)에 의해 크게 영향받지 않음.

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

중앙값(Median) & 최빈값(Mode)

- 최빈값: 데이터에서 가장 자주 나타나는 값
  - 측정 수준에 상관없이 사용될 수 있다는 점에서 가장 범용성이 높음.
  - 그러나 실질적으로 분석적 함의를 크게 가지고 있지 못함.
  - 대개 이항 변수일 경우에만 사용

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

분산(Variance): 범위(Range)와 분위(Centiles)

- 범위: 표본에서 최소값과 최대값 사이의 공간을 의미
- 분위: 특정한 값이 분포의 어디에 속하는지를 정량화하여 나타낸 결과
- 다른 측정지표들과는 달리, 범위와 분위는 모든 정보량에서 사용되지는 않음.
    - 예) 명목형 변수인 종교가 있다고 하자. 천주교, 기독교, 불교 등으로 코딩된 이 변수의 범위와 분위를 구할 수 있을까?

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

분산(Variance): 편차(Deviations)

- 편차($x_i - \bar{x}$)란 개별 값이 평균으로부터 떨어져 있는 단순 거리를 의미
- 분포에서 모든 편차의 총합은 0
- 모든 값의 편차를 제곱하여 그 평균을 구하면 표본의 분산(variance, σ_x)

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

분산(Variance): 편차(Deviations)

- 표준편차는 분산의 제곱근 값
    - 분산이 편차를 제곱하여 더한 것의 평균이었다면, 그러한 분산에 제곱근을 씌워줌으로써 원래의 측정 단위로 원상복귀시키는 것과 같음.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- n이 아니라 $n-1$로 나눠주는 이유는 Bessel의 제안에 따른 것

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

분산(Variance): 오차(Errors)

- 표본의 크기에 의해 가중치가 주어진(weighted) 표준 편차

$$\sigma_x = \frac{s}{\sqrt{n}}$$

- 표본평균의 정확성을 보여주는 신뢰구간(confidence intervals)을 계산할 때 사용

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

분산(Variance): 모먼트(Moment)

분포에 대한 일련의 속성들을 보다 일반적으로 보여줌.

- 어떤 분포의 K 번째(K_{th}) 모먼트 = M_K = $E[(x-\mu)^k]$
- $M_1 = E(x) = \bar{x}$ (평균 또는 중심경향성을 보여주는 다른 지표)
- $M_2 = E[(x-\mu)]^2 = \sigma^2$ (분산)
- $M_3 = E[(x-\mu)]^3$ = 왜도 (분포가 어떻게 기울어있는지)
  - 만약 $M_3 \geq 0$, 오른쪽으로 긴 꼬리를 가진 분포(right-skewed)
  - 만약 $M_3 \leq 0$, 왼쪽으로 긴 꼬리를 가진 분포(left-skewed)

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

분산(Variance): 모먼트(Moment)

- $M_4 = E[(x - \mu)^4]$ = 첨도 (kurtosis, 분포가 얼마나 뾰족한지)
  - lepto-: 매우 분포가 뾰족한 (한 쪽으로 집중되어 있는)
  - meso-: 분포가 비교적 정규형태로 잘 분포되어 있는
  - platy-: 분포가 상대적으로 평평한

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

히스토그램

R Code | Plot | Explanation

```
library(car);library(ggplot2);library(tidyverse)
Prestige %>% ggplot(aes(x = income)) +
  geom_histogram(color = "gray", binwidth = 2500) +
  scale_x_continuous(breaks = seq(0, 25000, 5000)) +
  labs(title = "Average Income in 1970 (dollars)") +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

커널 밀도 플롯

R Code | Plot | Explanation-1 | Explanation-2

```
Prestige %>% ggplot(aes(x = income)) +
  geom_histogram(aes(y = ..density.., color =),
                 breaks=seq(0,max(Prestige$income),2000)) +
  geom_density(aes(color = "Adaptive bandwith"),
                   kernel = "gaussian", linetype = 1, adjust = 0.8) +
  geom_density(aes(color = "Fixed bandwith"), bw = 1500, linetype = 2,
               position = "stack") +
  geom_rug(color = "grey") +
  scale_x_continuous(breaks = c(seq(0, 30000, 5000))) +
  scale_y_continuous(breaks = c(seq(0, 15e-5, 5e-5))) +
  theme_bw() + theme(legend.title = element_blank(), legend.text = element_text(size = 6),
                     axis.text.y = element_text(angle = 90, hjust = 0.5),
                     legend.position = c(0.75, 0.75))
```

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

Q-Q Plot

R Code | Plot | Explanation

```
with(Prestige, {
  par(mfrow = c(1, 2), mar = c(4, 4, 4, 4))
  qqPlot(income, id=list(n=0), col.lines="black")
  plot(density(Prestige$income), main = "")})
```

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

박스플롯(Boxplots)

R Code | Plot | Explanation

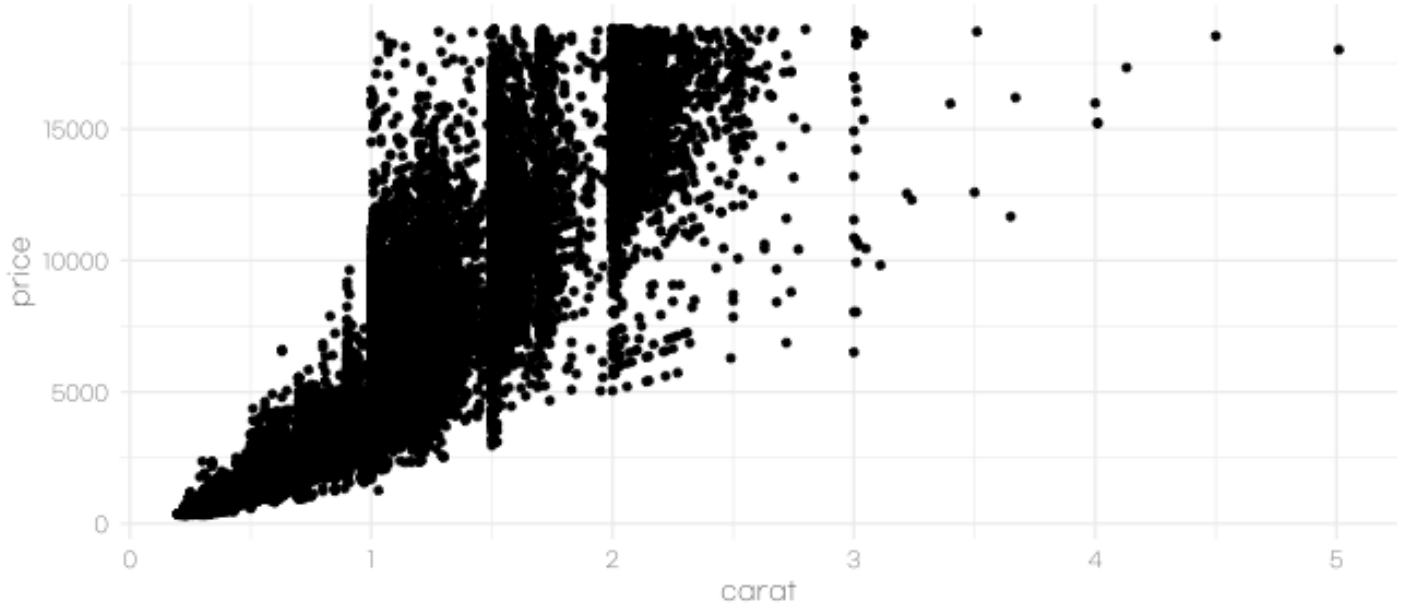
```
is_outlier <- function(x) {
   return(x < quantile(x, 0.25) - 1.5 * IQR(x) |
             x > quantile(x, 0.75) + 1.5 * IQR(x))}
Prestige %>% rownames_to_column() %>%
   mutate(outliers = ifelse(is_outlier(income),
                            rowname, NA_character_)) %>%
   ggplot(aes(x = "", y = income)) + geom_boxplot() +
   ggrepel::geom_text_repel(aes(label=outliers)) + theme_bw() +
   labs(x = "")
```

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

산포도(Scatterplots)

Plot Explanation



연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

산포도(Scatterplots)

R Code | Plot

```
library(patchwork)
p1 <- Ornstein %>% ggplot(aes(x = assets)) + geom_density() +
  labs(x = "assets", subtitle = "(a)") + geom_rug() + theme_bw()
p2 <- Ornstein %>% ggplot(aes(x = log10(assets))) + geom_density() +
  scale_x_continuous(labels = c(100, 1000, 10000, 100000)) +
  labs(x = latex2exp::TeX("log$_{10}$(assets)"), y = "", subtitle = "(b)") +
  geom_rug() + theme_bw()
p3 <- p1 + p2 + plot_layout(ncol = 2)
print(p3)
```

연구방법의 적용 III. 중심경향성 및 관계 파악

중심경향성의 측정

산포도(Scatterplots)

R Code-1 R Code-2 Plot

이산형 데이터의 경우, 관측치들이 중첩되어 퍼져있는 정도를 파악하기 어려울 수 있는데, 이 경우 `jitter`로 해결.

```
library(ggExtra)
p5 <- Vocab %>% ggplot(aes(vocabulary, education)) +
  geom_point(size = 2, alpha = 0.5) +
  geom_smooth(method = "lm", se = T, color = "black") + theme_bw()
p5 <- ggMarginal(p5, type="boxplot", size = 15)
```

연구방법의 적용 III. 중심경향성 및 관계 파악

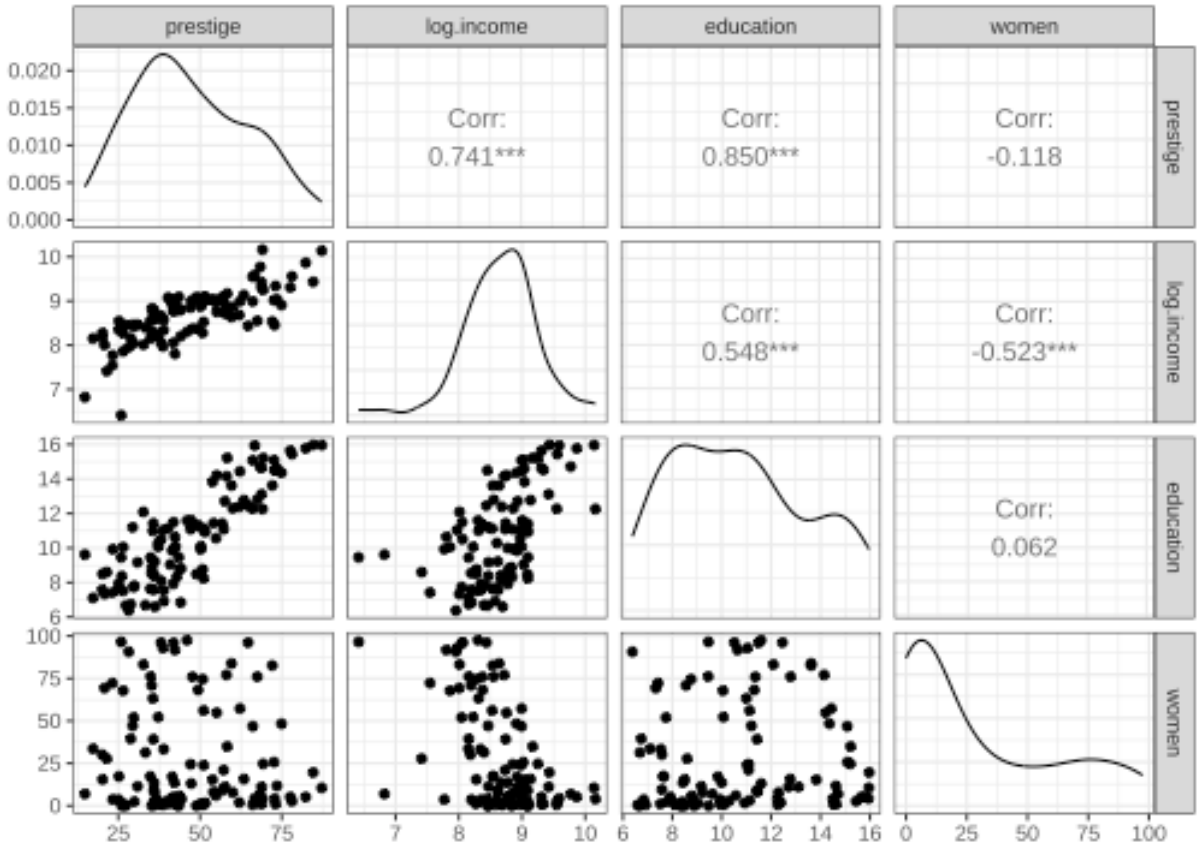
데이터 변환

왜 데이터를 변환(transform) 해야할까?

- 한쪽으로 치우친 분포를 가지고 있는 데이터는 많은 관측치들이 좁은 범위에 모여있기 때문에 분석이 어려움.
- 치우친 분포에서는 대개 비정상적으로 큰 값과 작은 값들이 나머지 값들을 살펴보기 어렵도록 짓누름(suppress).
- 분포를 요약해서 보여주는 통상의 통계 방법들은 대개 평균을 이용
    - 문제는 평균이 극단적인 값, 이탈치들에 민감한 통계치
    - 따라서 서로 다른 척도/범위를 가진 변수들을 비교하기란 어려울 수 있음.

연구방법의 적용 III. 중심경향성 및 관계 파악

데이터 변환



연구방법의 적용 III. 중심경향성 및 관계 파악

데이터 변환

원칙 | Bulging Rule | R Code-1 | R Code-2 | Plot

거듭곱 변환(Power transformations)도 관계를 명확하게 보여주는 데 도움이 될 수 있음.

> 단순한 비선형관계는 종종 X, Y 또는 둘 모두를 거듭곱 변환을 함으로써 바로잡을 수 있다. Mosteller와 Tukey의 's bulging rule은 선형화 변환을 선택하는 데 도움을 준다.

연구방법의 적용 III. 중심경향성 및 관계 파악

데이터 변환

정규화(Normalization)

마지막으로 살펴볼 데이터 변환: 정규화 & 표준화

- 둘 모두 서로 다른 척도/단위의 변수를 동일한 척도로 변환하여 비교할 수 있게 해줌.
- 정규화: [0, 1] 사이의 범주로 데이터를 변환
  - 각 값에서 최소값을 뺀 이후에 최대값에서 최소값을 뺀 값으로 나누어줌.
  - 최소최대값이 계산에 반영되는 정규화는 이탈치에 민감
  - 대개 머신러닝에서는 사용하지만, 통계모델에서는 사용하지 않음.
  - $\text{Normalization} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$

연구방법의 적용 III. 중심경향성 및 관계 파악

데이터 변환

표준화(Standarization)

표준화(또는 z-스코어 정규화)는 변수를 0으로 중심화(centering)하고 분산을 1로 표준화한다는 것을 의미

- 표준화: 각 관측치들로부터 평균을 빼고 그 값들을 표준편차로 나눔
  - 다른 척도를 가진 변수들이 동일한 표준정규분포의 특성을 가지도록 함.
  - 표준화 결과로 나타나는 최소값과 최대값은 변수가 어떻게 퍼져있는지에 따라서 다르고 이탈치(outliers)의 존재 여부에 매우 크게 영향을 받음.
  - $\text{Standardization} = \frac{x - \bar{x}}{s}$

KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

감사합니다!

궁금한 것이 있으면 언제든 연락하세요.

강사 연락처

연락처	박상훈
	sh.park.poli@gmail.com
	sanghoon-park.com/
	영상바이오관 405