L04 - Descriptive Statistics

Sim, Min Kyu, Ph.D., mksim@seoultech.ac.kr

SEOULTECH 국립서울과학기술대학교 IT정책전문대학원

Motivation

- 어떤 농구선수의 최근 20경기 리바운드 기록이다.

```
x <- c(4, 5, 5, 6, 6, 7, 8, 8, 8, 8,
       9, 9, 10, 11, 12, 12, 12, 13, 13, 14)
```

- 우리는 이를 어떻게 요약하여 이해하는가?

Central tendency (A measure of location)

- 평균 (mean)
  - 가장 널리 쓰이는 대표값 `mean(x)`
  - 수치들의 합을 관찰값의 갯수로 나눈 것 `sum(x)/length(x)`
- 중간값 (median)
  - 자료값을 정렬하였을 때의 가운데 값 `median(x)`
  - 중간값 이상인 경우가 반, 중간값 이하인 경우가 반으로 나뉜다.
  - 잘하지도 못하지도 않은 경기의 리바운드 횟수
- 최빈값 (mode)
  - 가장 잦은 관찰값 (누가봐도 4회나온 8)
- 평균, 중간값, 최빈값이 일상생활에서 부정확하게 사용되는 경우가 매우 많다.

```
sum(x)
```

```
## [1] 180
```

```
length(x)
```

```
## [1] 20
```

```
mean(x)
```

```
## [1] 9
```

```
median(x)
```

```
## [1] 8.5
```

- 다음 함수는 이러한 기초적인 내용을 한번에 출력해준다.

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    6.75    8.50    9.00   12.00   14.00
```

- `Min.`: minimum (최소값)
- `Max.`: maximum (최대값)
- `1st Qu.` (Q1): first quartile (하위 25%)
  - 75%가 이 수치보다 크다.
- `3rd Qu.` (Q2): third quartile (상위 25%)
  - 75%가 이 수치보다 적다.

Motivation

- 두 선수 X와 Y가 있다. 이들 두 선수의 기록이 각각 아래와 같다.

```
x
```

```
##  [1]  4  5  5  6  6  7  8  8  8  8  9  9 10 11 12 12 12 13 13 14
```

```
y <- c(rep(8,5), rep(9,10), rep(10,5))
y
```

```
##  [1]  8  8  8  8  8  9  9  9  9  9  9  9  9  9 10 10 10 10 10
```

- 두 선수의 리바운드 평균을 비교해보면 다음과 같다.

```
mean(x)
```

```
## [1] 9
```

```
mean(y)
```

```
## [1] 9
```

- 당신이 감독이면 누가 더 믿을만한 선수이며, 이를 어떻게 표현할 수 있는가?

Measures of dispersion

1. Range (maximum-minimum)를 사용할 수 있다.

```
max(x)-min(x)
```

```
## [1] 10
```

```
max(y)-min(y)
```

```
## [1] 2
```

2. IQR(inter-quartile range)을 사용할 수 있다.

```
as.numeric(summary(x)[5]-summary(x)[2])
```

```
## [1] 5.25
```

```
as.numeric(summary(y)[5]-summary(y)[2])
```

```
## [1] 0.5
```

- 분산/표준편차를 사용할 수 있다?

분산과 표준편차

편차 (deviation)

- 각 관찰값과 평균과의 차이를 **편차 (deviation)**라고 한다.

```
x - mean(x)
```

```
##  [1] -5 -4 -4 -3 -3 -2 -1 -1 -1 -1  0  0  1  2  3  3  3  4  4  5
```

```
y - mean(y)
```

```
##  [1] -1 -1 -1 -1 -1  0  0  0  0  0  0  0  0  0  0  1  1  1  1  1
```

- 편차를 어떻게 이용할 것인가?
  - 편차의 평균을 내면 항상 0이다.
  - 편차를 고려할때에 더이상 부호(sign)은 중요하지 않으므로 부호를 제거하는 과정이 필요하다.

3. 평균 절대편차 (MAD: Mean Absolute Deviation)

- 편차의 절대값을 씌워서 부호를 제거하고 평균을 낸다.

```
mean(abs(x-mean(x)))
```

```
## [1] 2.5
```

```
mean(abs(y-mean(y)))
```

```
## [1] 0.5
```

4. 분산 (Variance)

- 편차를 제곱하여 (부호가 제거되며) 평균을 낸다.
- 평균에서 많이 벗어나면 **제곱**만큼이나 큰 영향을 준다.

```
mean((x-mean(x))^2) # sum((x-mean(x))^2)/length(x) # var(x)
```

```
## [1] 8.6
```

```
mean((y-mean(y))^2) # sum((x-mean(x))^2)/length(x) # var(y)
```

```
## [1] 0.5
```

- MAD와 Variance의 선택은 가치판단을 요구한다.
  - a와 b의 deviation이 아래와 같다면…
  - MAD는 같다. 그러다 Variance는 b의 경우가 더 크다.
  - 일반적인 경우에는 Variance를 쓴다.

```
dev_a <- c(-1, -1, 1, 1)
dev_b <- c(-2, 0, 0, 2)
c(mean(abs(dev_a)), mean(abs(dev_b))) # MAD
```

```
## [1] 1 1
```

```
c(mean(dev_a^2), mean(dev_b^2)) # Variance
```

```
## [1] 1 2
```

- MAD와 Variance의 단위
  - MAD의 단위는 원래 수치의 단위인 “리바운드 갯수”
  - Variance의 단위는 원래 수치 단위의 “제곱”인 “리바운드 갯수의 제곱?”
  - Variance의 단위를 원래 수치로 맞춰주기 위해서 제곱근을 취한다. 이 수치를 **표준편차 (standard deviation)**이라고 한다.

5. 표준편차 (Standard Deviation, SD, sd, Std. Dev.)

- Variance의 제곱근으로 정의된다.

```
sqrt(mean((x-mean(x))^2)) # sd(x) or sqrt(8.6)
```

```
## [1] 2.932576
```

```
sqrt(mean((y-mean(y))^2)) # sd(y) or sqrt(0.5)
```

```
## [1] 0.7071068
```

6. 변동계수 (Coefficient of Variation, CV)

- $CV = —$
- [평균 1, 표준편차 1] vs [평균 10, 표준편차 1]인 경우는 편차의 차이가 있다.
- 표준편차와 평균은 단위가 같으므로 이들의 비율을 측정한다.

Summary for “Measure for dispersion”

1. Range (maximim-minimum)
2. IQR (Q3-Q1)
3. MAD
4. Variance
5. Standard Deviation
6. Coefficient of Variation

모집단(Population)과 표본(Sample)

- 이번 사례의 분석의 목적은 무엇인가? 우리가 알고 싶은것은
  - 해당 농구선수의 본질적인 능력과 향후 performance에 대한 예측이지
  - 20경기에 대한 진실만을 알려고 하는 것이 아니다.
- 그러므로 우리의 관심이 되는 모수(parameter)는 **해당 선수의 경기당 리바운드 갯수**이다.
- 표본집단의 통계량을 바탕으로 아래와 같이 **추정(estimation)**한다.
- 이렇게 추정을 위해 사용하는 통계량을 **추정량(estimator)**라고 한다.

모수 추정의 예시

1. 최근 20개 경기의 리바운드 갯수에서
   - `mean(x)`의 값이 9이므로,
   - **해당 선수의 경기당 리바운드 갯수**의 평균은 9일 것으로 추정된다.
2. 최근 20개 경기의 리바운드 갯수에서
   - `median(x)`의 값이 8.5이므로,
   - **해당 선수의 경기당 리바운드 갯수**의 중간값은 8.5일 것으로 추정된다.
3. 최근 20개 경기의 리바운드 갯수에서
   - `sum((x-mean(x))^2)/length(x)`의값이 8.6이므로,
   - **해당 선수의 경기당 리바운드 갯수**의 분산은 8.6일 것으로 추정된다??
   - 그렇지 않다!