통계학 입문

2021-03-21

데이터과학융합스쿨

손 대 순 교수

앞서 배운 내용을 가지고 몇 가지 실습을 한 번 해 볼까요?

R studio를 켜고 아래의 내용을 한 번 따라해 보세요.

```
> a <- sample(c(10:20, 30:50), 100, replace=T)
> a
  [1] 46 18 18 44 15 39 31 38 16 16 45 33 19 20 16 37 46 14 35 32 15 37 10 50 10 18 33 32
 [29] 20 44 37 46 49 38 47 50 14 37 12 19 33 50 38 18 18 10 15 19 48 35 40 39 15 17 19 20
 [57] 30 18 45 12 14 48 49 12 45 10 10 42 15 20 12 18 42 45 35 41 14 18 13 12 10 30 41 50
 [85] 38 47 34 35 33 11 37 46 48 37 37 41 41 46 36 47
> mean(a)
[1] 29.95
> median(a)
[1] 33
> mode(a)
[1] "numeric"
> which.max(a)
[1] 24
> table(a)
a
10 11 12 13 14 15 16 17 18 19 20 30 31 32 33 34 35 36 37 38 39 40 41 42 44 45 46 47 48
 6  1  5  1  4  5  3  1  8  4  4  2  1  2  4  1  4  1  7  4  2  1  4  2  2  4  5  3  3
49 50
 2  4
> which.max(table(a))
18
 9
> var(a)
[1] 183.5833
> sd(a)
[1] 13.54929
> sd(a)/mean(a)*100
[1] 45.23971
> range(a)
[1] 10 50
> diff(range(a))
[1] 40
```

첫 줄부터 한번 살펴 봅시다.

sample이라는 함수는 주어진 범위에서 숫자를 랜덤하게 추출하는 함수죠. 이번에는 조금 응용해서, 주어진 범위를 10~20, 30~50 이렇게 주었습니다. 그러니, 0~9나 21~29까지의 숫자는 뽑히지 않겠죠? 총 100개의 숫자를 뽑고, 중복은 허용합니다. 변수 a라고 하는 곳에 뽑힌 숫자들이 저장 됩니다.

a에 무엇이 들어 있는지 한 번 살펴 볼까요? 그냥 a만 쳐 주면 됩니다. 주어진 범위 내에서

100개의 숫자가 추출되었는지 한 번 보세요. 평균, 중앙값은 그림에서 보여지는 대로 mean이나 median 함수를 사용하면 됩니다. 최빈값을 구하기 위해서는 mode를 사용하면 될까요? 보는 것처럼 이상한 값이 나오죠? R에서 mode함수는 변수의 성격을 보여주는 함수입니다. 이것이 숫자로 이루어진 변수인지, 문자로 이루어진 변수인지를 나타내 주는 것이지요. 그럼 최빈값을 구하기 위해서는 어떻게 해야 할까요?

R에서 which.max라는 함수가 있어요. 최대값이 어디 있는지를 찾아주는 함수예요. which.max(a)를 입력하니, 24라는 값이 나왔네요. 여러분들은 다를 수 있겠지요. 추출된 값이 다를 테니. 위 그림에서 24라는 값의 의미는 24번째 숫자가 max값이라는 뜻이예요. 그림에서 보면 24번째 값이 50이고 이 값이 max라는 의미가 되는거죠. 그렇다면, 이 which.max도 우리가 알고 있는 최빈값을 구하는 방법은 아니네요. 그렇죠?

표를 만들어 봅시다. table(a)라는 명령으로, 우리가 뽑은 숫자들이 어떻게 분포하고 있는지를 살펴볼 수 있어요. 위 그림에서 출력된 표를 보면, 주어진 범위에서 각 숫자들이 몇 개씩 뽑혔는지를 알 수 있지요. 이 테이블에서 which.max함수를 써 보면 어떨까요?

which.max(table(a))

라고 명령을 주는 거예요. 그러면 18과 9라는 결과가 나옵니다. 가장 빈도가 높은 수는 18이라는 뜻이고, 앞에서부터 9번째 자리에 있다는 말이예요. 결국 최빈값은 18이 되겠지요. 몇가지 단계를 거쳐야 하니 어려워 보이지만, 프로그램을 길게 짜다 보면, 이러한 방식이 매우 유용하답니다.

분산, 표준편차도, var나 sd 명령으로 구할 수 있어요. CV도 직접 구하는 함수를 제공하지는 않지만, 어렵지 않은 수식이니, 위 그림에서처럼 식을 입력하는 방법으로 구할 수 있지요. 또 range라는 함수는 최소값과 최대값을 보여주는 함수예요. 실제 통계량으로 사용하는 range는 최대값과 최소값의 차이니까, diff라는 함수를 같이 사용해 주면 금방 구할 수 있겠지요.

자. 지금부터는 여러분들 스스로 한 번 해 보세요.

quantile라는 함수로 분위부, 사분위수도 구할 수 있습니다. 하지만 간단하게 되지는 않을거예요. 함수를 사용하는 문법은 ?quantile과 같이 함수 이름 앞에 ?를 주면 사용설명서가 옆에 뜰거예요.

왜도 첨도도 구해 봅시다.

skewness(a), kurtosis(a)와 같이 명령어를 주면 구할 수 있지만, 이는 R 기본 함수로 제공하지 않기 때문에 에러 메시지가 뜰겁니다. 이런 경우에 필요한 package를 설치해 주어야 합니다.

왜도나 첨도를 구할 수 있는 패키지는 "fBasics"라는 패키지에서 제공하고 있어요.

이를 설치하는 방법은

install.packages("fBasics")

라고 입력해 주면 됩니다. 그러면 쭈우욱~ 설치가 되는 화면이 나올거예요.

설치가 완료되면

library(fBasics)

라고 입력합니다. 이는 해당 패키지를 사용하겠다는 선언이지요. 아무런 메시지가 나오지 않는 것이 정상입니다.

그리고 나서 skewness(a), kurtosis(a)와 같은 명령어를 입력해 보세요.

결과가 나오나요?

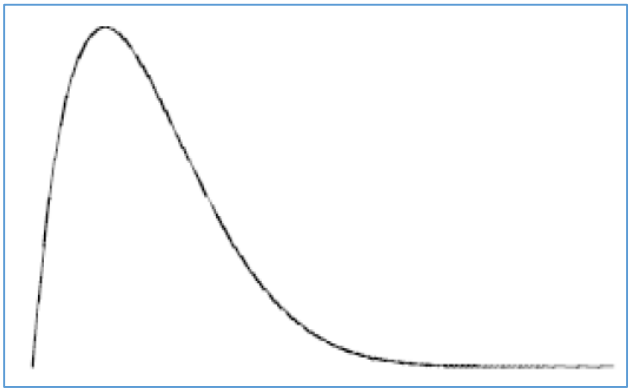
이와 같이 R에서 패키지를 사용하는 방법을 여러분들 스스로 한 번 해보는 경험을 가져 봤습니다. quantile과 왜도, 첨도를 구하는 과정을 직접 수행해 보고, 인증샷을 과제로 제출해 봅시다.

조금 덜 친절한 것에 대해서도 여러분들이 차츰 적응해야지요? ㅎㅎ

여러분들의 필요에 따라서, 함수를 찾아보고, 패키지도 설치해서 결과를 내는 방법에 익숙해 져야 합니다.

앞 시간에 예로 들었던 내용을 다시 한 번 살펴보고, R로 실습도 해 봅시다.

4명의 직원에게 10원의 월급을 주고 사장은 10억의 월급을 받는 악덕 기업에 대해 이야기 했었지요. 숫자가 적기는 하지만, 월급을 적게 받는 사람이 많고, 많이 받는 사람이 적은 분포입니다. 대략 분포를 그려보면, 아래와 같겠지요.



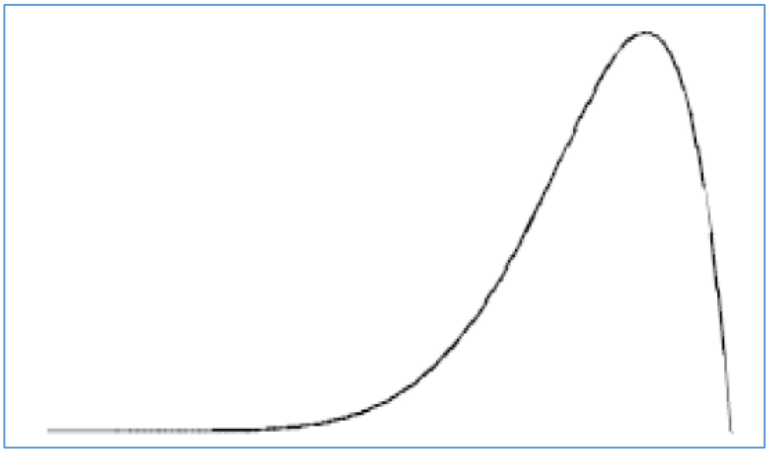
왼쪽으로 치우쳐진 분포입니다. Right skewed distribution이죠. 앞에서 배운 것들이 막 떠올라야 합니다. ^^ skewness 통계량을 계산하면 양수(+)가 나오겠네요. R에서 한 번 해 보세요.

library(fBasics)라는 명령어는 앞에서 수행했다면 또 해주지 않아도 됩니다. R 혹은 Rstudio를 끄지 않으면 메모리에 불러져 있는 상태(사용 가능한 상태)이니까요.

```
> library(fBasics)
Loading required package: timeDate
Loading required package: timeSeries
> a <- c(10, 10, 10, 10, 1000000000)
> skewness(a)
[1] 1.073313
attr(,"method")
[1] "moment"
>
```

자. 맞죠? 이제 여러분이 이런 것들을 직접 해 보면서 확인할 수 있어야 합니다.

자, 다시 회사 얘기로 돌아가서, 이 회사의 중앙값은 얼마인가요? 10이죠. 5개의 값 중 3번째 값이니까요. 그럼 평균은? 굳이 정확하게 계산할 필요는 없습니다. 대략 10억/5 하면 2억쯤 되겠네요. 평균이 중앙값보다 엄청 크다는 이야기입니다. 사장이 가지고 가는 월급이 10억, 100억, 1000억 늘어날 때마다 평균인 함께 커지게 되지요. 분포의 모양으로 말하면, 꼬리가 길어질수록 평균이 꼬리쪽으로 움직이게 되어 있습니다. 꼬리가 평균을 데리고 다니는 것이죠. 이를 정리하면, 평균과 중앙값을 비교하여 "평균>중앙값"의 관계가 있으면, 이는 왼쪽으로 치우친 분포 (right skewed distribution)라는 의미가 됩니다. 1,2,3,4,5와 같이 3을 중심으로 좌우 대칭인 데이터라면 "평균=중앙값"의 관계가 되겠지요.



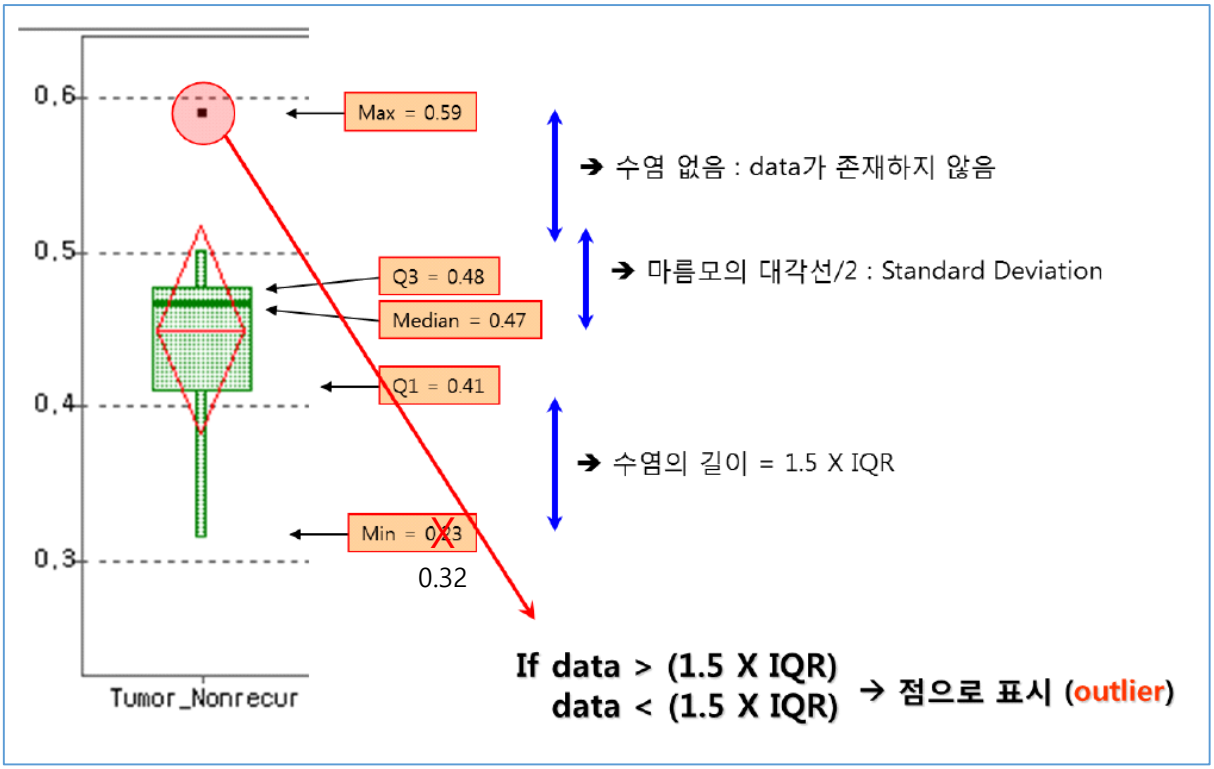
위와 같은 분포를 생각해 봅시다. 오른쪽으로 치우쳐진 분포예요. 회사 월급을 예로 든다면, 월급을 10원 받는 직원이 하나 있고, 나머지 4명은 10억을 받는다고 합시다. 이런 회사를 알고 있으면 나에게 소개 좀 시켜 주세요. ^^ 10억을 받는 직원이 되어야 할텐데 말이죠... ㅎㅎ

이 회사의 월급 평균과 중앙값을 구해 보죠. 중앙값은 10억인데 반해, 평균은 중앙값보다 한참 낮죠? 약 8천만원 정도 되겠네요. "중앙값>평균"의 관계입니다. 평균이 중앙값보다 작다는 것은 평균 쪽(작은 쪽)으로 꼬리가 형성되어 있는 것이예요. 그러니 left skewed distribution이고, 오른쪽으로 치우쳐진 분포이며, skewness값은 음수가 되겠네요. 어때요? 생각할 것이 많지요?

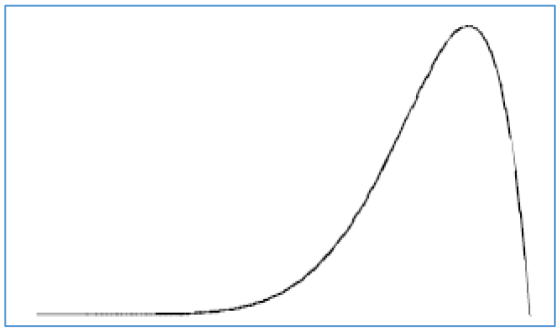
왼쪽으로 꼬리가 길면 길수록 평균을 많이 끌고 내려옵니다. 오른쪽으로 꼬리가 있다면, 평균을

끌고 올라가는 형태가 되는 것이지요. 평균과 중앙값의 숫자만 가지고도 대략의 분포 모양이 머리 속에 그려져야 합니다.

여러가지를 한꺼번에 생각해야 하니, 머리가 조금 복잡하긴 하겠지만, 위의 내용을 모두 이해하고 난 다음에 다음의 boxplot을 한 번 다시 봅시다.



자, 이제 많은 정보들이 보이지요? 평균은 0.45정도 되는 것 같구요. 중앙값은 0.47로 표시되어 있습니다. 평균보다 중앙값이 더 크군요. 이것은 값이 큰 방향으로 치우쳐져 있음을 의미합니다. 왼쪽과 같은 분포 형태를 가진 것이지요.



Boxplot의 빨간색 마름모가 없다고 합시다. 이는 그림에서 평균이 어디인지 정확히 모른다는 뜻이예요. 그렇다면 분포가 왼쪽 그림과 같다는 것을 알 수 없을까요? 그래도 짐작할 수 있습니다.

위 boxplot의 박스 내부를 들여다 봅시다. 중앙값이 박스 위쪽에 바짝 붙어 있지요? 아래쪽 수염에 비해 위쪽 수염의 길이도 짧습니다. 이것이 무엇을 의미할까요? 전체 데이터가 100개 있다

면 (min~Q1), (Q1~Q2), (Q2~Q3), (Q3~max) 이 네 개의 구간에 모두 데이터가 25개씩 들어 있다는 말입니다. 그런데 (Q1~Q2)의 길이에 비해 (Q2~Q3)의 길이가 짧다? 이는 짧은 구간 내에 25개의 데이터가 밀집해 있다는 것을 의미하지요. (min~Q1)이나 (Q3~max) 구간도 마찬가지로 전체 데이터의 25%를 가지고 있는 구간입니다. 그런데 구간의 길이가 상대적으로 짧다면, 그만큼 데이터가 밀집해 있다는 뜻이겠지요. 사람으로 치면 인구밀도가 높은 것입니다.

자, 여러분, 고생했어요. 박스그림 이야기만 가지고 한참 했네요. 손목이 아파 더 치기 어렵습니다. ^^ 아무튼 이 boxplot이 가지고 있는 정보량이 이렇게 많습니다. 이 그림 하나로, 전체적인 분포 모양이 머리 속에 떠 올라야 하고, 그림에서 보이지 않더라도, 평균/중앙값의 관계가 떠 올라야 합니다.

R에서 boxplot 그리는 명령어는 그냥 'boxplot'입니다. ^^

```
> a<-sample(1:100, 500, replace=T)
> boxplot(a)
> b<-c(a,200)
> boxplot(b)
> c<-c(-10,b)
> boxplot(c)
> d<-c(-30,b)
> boxplot(d)
> e<-c(-50,b)
> boxplot(e)
> e<-c(-70,b)
> boxplot(e)
> ?boxplot
>
```

자, 이런 명령어 이제 좀 이해할 수 있을까요? 조금씩 응용해 가는 과정입니다. 무엇을 뜻하는 것인지 그려보면서 확인해 봅시다. 각 변수들에 무엇이 들어 있는지 확인해 보면, 금방 알 수 있어요. 다양한 응용들을 해 보아야 합니다.

이제 인증샷을 찍어 올리라는 과제를 주지 않아도 여러분들 스스로 해 볼 수 있죠? 과제는 9강의 내용까지만으로 하죠. boxplot 등 11강에서 이야기한 것을 함께 올려 주어도 좋습니다. 필수는 아니구요. 고등학생처럼 꼭 숙제 검사를 해야만 하는 습관...ㅋㅋ 이제 안 그러기로 했잖아요.

다음 시간부터는 확률 이야기를 해야겠습니다. 확률은 통계의 언어라고 하지요. 이제 몸을 좀 풀었으니, 한 번 놀아 봅시다... ^^