범주형 자료 표현하는 방법(categorical data)

범주형 자료의 요약

범주형 자료에서는 각 관측 값의 크기가 아니라 자료가 갖는 범주의 종류에 관심이 있으므로, 각 범주가 나타나는 **횟수**를 요약함으로써 범주형 자료의 개요를 파악할 수 있다.

---

1. 도수분포표(frequency table)

(1) 이산형 도수분포표

범주형 자료의 경우 각 관측값은 몇 개의 범주 중 하나의 값을 갖게 된다. **도수**(frequency)는 각 범주에 속하는 관측값의 개수를 말한다. 도수를 전체 개수로 나눈 비율(전체 값에서 차지하는 상대적인 비중)을 그 범주의 **상대도수**(relative frequency)라고 한다[1]. **도수분포표**(frequency table)는 범주형 자료에서 범주(계급), 그 범주에 대응하는 도수, 상대도수를 나열하여 표로 작성한 것을 말한다.

[예제] 한 회사에서 새로 개발한 자동차의 외형에 대하여 고객 150명을 임의로 뽑아 선호도를 조사하였다. 150명 중에서 **71명**은 좋다고 답하고, **42명**은 보통이다, 28명은 싫다, **9명**은 답을 하지 않았다. 이 조사에 대한 도수분포표를 작성하여라.

답	도수	상대도수
좋다	71	0.473(71/150)
보통이다	42	0.280
싫다	28	0.187
무응답	9	0.060
합	**150**	**1.000**

예제에 있는 테이블을 R에서 구현하면 다음과 같다.

```
data <- data.frame("답"=c("좋다","보통이다","싫다","무응답","합"),
                   "도수"=c(71,42,28,9,150))
data
```

1) 상대도수를 추가하여 전체 값에서 상대적인 비중을 차지하는 값을 비교할 수 있다.

```
data$상대도수 <- round(data$도수/150, 3)  # 3은 반올림하여 소숫점 3째 자리까지
data
```

```
          답  도수  상대도수
1        좋다   71     0.473
2    보통이다   42     0.280
3        싫다   28     0.187
4      무응답    9     0.060
5          합  150     1.000
```

다른 통계 패키지나 언어에서도 반올림은 널리 사용되는 방식이다. 소수점 아래까지 너무 길게 나오는 것을 적당히 줄여서 해석하기 쉽게 만드는 것이다.

반올림 함수 round( )의 사용법

round(개체, digits=소수점 아래 반올림 하여 나타내고 싶은 위치)

```
data <-data.frame("답"=c("좋다","보통이다","싫다","무응답","합"),
                  "도수"=c(71,42,28,9,150))
상대도수 <- round(data$도수/150, 3)
data <- cbind(data, 상대도수)
data
```

R의 table( ), xtabs( ) 함수를 이용하여 **도수분포표, 상대도수분포표**를 작성하자.

[예제] MASS 패키지에 내장된 Cars93 데이터프레임의 차종형태(Type) 변수를 분석해 보자.

```
library(MASS)
str(Cars93)
View(Cars93)
```

'data.frame': 93 obs. of 27 variables:
 $ Manufacturer : Factor w/ 32 levels "Acura","Audi",..: 1 1 2 2 3 4 4 4 4 5 ...

```
 $ Model     : Factor w/ 93 levels "100","190E","240",..: 49 56 9 1 6 24 54
 $ Type      : Factor w/ 6 levels "Compact", "Large",.. : 4 3 1 3 3 3 2 2 3 2
 $ Min.Price : num  12.9 29.2 25.9 30.8 23.7 14.2 19.9 22.6 26.3 33 ...
 $ Price     : num  15.9 33.9 29.1 37.7 30 15.7 20.8 23.7 26.3 34.7 ...
 $ Max.Price : num  18.8 38.7 32.3 44.6 36.2 17.3 21.7 24.9 26.3 36.3 ...
```

(1) 도수분포표(frequency distribution table) : table(), xtabs()

① table( ) 함수로 단순한 표의 형태로 표현할 수 있다.

```
freq <- table(Cars93$Type) ; freq # 차종형태 변수의 유형별 빈도수
```

```
Compact   Large Midsize   Small  Sporty     Van
     16      11      22      21      14       9
```

② prop.table( ) 함수를 이용하여 상대도수를 구할 수 있다.

```
propfreq <- prop.table(freq) ; propfreq  # 상대도수분포표
propfreq <- round(propfreq, 3) ; propfreq  # round(개체, digits=소수점 아래 반올림하고 싶은 위치)
```

③ rbind( ) 함수를 이용하면 도수(freq) 행과 상대도수 행을 결합할 수 있다.

```
table <- rbind(freq, propfreq) ; table     # 도수와 상대도수의 행 결합
```

```
         Compact  Large Midsize  Small  Sporty   Van
freq      16.000 11.000  22.000 21.000 14.000 9.000
propfreq   0.172  0.118   0.237  0.226  0.151 0.097
```

[실습1] 위의 결과를 **전치행렬** 형태로 나타내는 R-코드를 작성하여라.

④ addmargins( ) 함수를 이용하여 합을 추가한다.

addmargins( ) 함수: 행과 열의 합을 계산하기 위한 함수

```
addmargins(table)      # 행과 열의 합
table <- addmargins(table, margin=2) ; table    # 1: 열의 합 2: 행의 합
```

	Compact	Large	Midsize	Small	Sporty	Van	Sum
freq	16.000	11.000	22.000	21.000	14.000	9.000	93.000
propfreq	0.172	0.118	0.237	0.226	0.151	0.097	1.001

[실습2] [실습1]에서 도수의 합과 상대도수의 합을 나타내는 R-코드를 작성하여라.

(2) 연속형 자료의 도수분포표

몸무게, 키, 성적 등 연속형 자료인 경우에는 **구간(계급)**을 정해 그 구간 안에 속한 자료의 개수를 표시한다.

예를 들어, 성적의 구간별 인원수의 도수분포에 **상대도수(비율), 누적상대도수**를 추가하여 40점대 또는 30점대 하위권 인원의 비율을 쉽게 알 수 있고, 90점대 이상 상위권 학생의 비율도 알 수 있다.

[예제] MASS 패키지에 내장된 Cars93 데이터프레임의 Price(차량가격) 변수를 분석해 보자.

```
library(MASS)
str(Cars93)
str(Cars93$Price)
print(Cars93$Price)
summary(Cars93$Price)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7.40	12.20	17.70	19.51	23.30	61.90

① 연속형 자료는 도수분포표를 만들기 위해서 구간을 나누어야 하며 이를 **계급**이라고 한다.

cut(자료를 구간으로 나눌 객체, breaks=계급의 개수, labels=계급의 이름)

```
cut(Cars93$Price, breaks=7)
7 Levels: (7.35,15.2] (15.2,23] (23,30.8] ... (54.1,62]
```

구간의 내용이 **(7.35,15.2]** 로 표시되는데 (은 포함시키지 않으며,]은 포함된다는 의미이다.

② **이산형 도수분포표**를 작성하는 과정과 같다.

```
cut(Cars93$Price, breaks=7)
Price <- cut(Cars93$Price, breaks=7)
freq <- table(Price) ; freq        # 차량가격 변수의 구간별 빈도수
propfreq <- prop.table(freq) ; propfreq   # 상대도수 분포표
propfreq <- round(propfreq, 3) ; propfreq   # 소수점 아래 반올림 하고 싶은 위치
table <- rbind(freq, propfreq) ; table      # 도수와 상대도수의 행 결합
addmargins(table)      # 행과 열의 합(여기서는 열의 합은 의미가 없다.)
table <- addmargins(table, margin=2) ; table
```

③ 연속형 자료이므로 **누적도수** 또는 **누적상대도수**는 **cumsum( )** 함수를 이용한다.

```
cumfreq<- cumsum(freq) ; cumfreq   # 누적도수
cumpropfreq<- cumsum(propfreq) ; cumpropfreq   # 누적상대도수
```

[실습3] (1) 누적도수와 누적상대도수를 행결합하는 R-코드를 작성하여라.

(2) 도수, 상대도수, 누적도수, 누적상대도수의 행 결합을 나타내는 R-코드를 작성하여라.

④ 위의 예에서는 계급의 개수를 입력하여 구간을 나눴지만, 구간은 직접 **break 포인트**를 설정하여 임의로 설정할 수 있다.

```
cut (개체, breaks = c(1, 10, 20, 30, 40, 50, 60, 70) )
cut( Cars93$Price, breaks = c(1, 10, 20, 30, 40, 50, 60, 70) )
table( cut(Cars93$Price, breaks = c(1, 10, 20, 30, 40, 50, 60, 70)) )
```