2018년 07월 04일 8일차

확률변수와 확률분포

1) 조건부 확률

> 사전 정보값이 주어지며, 해당 정보의 확률 내에서 특정 사건이 일어날 확률을 계산한다.

베이지안 통계학 / 베이즈 정리
나이브 - 베이즈 정리 : 범주형 / 이산형 데이터 분석에 사용

```
Pxy <- matrix(c(0, 1/4, 1/4, 1/4, 1/4, 0), ncol = 3, byrow = T)
dimnames(Pxy) <- list(0:1, 0:2)
Py <- margin.table(Pxy, margin = 1)
Px <- margin.table(Pxy, margin = 2)
Pxy[2,]/Py[2]
```

2) 조건부 확률밀도함수

```
fx1 <- function(x) { (2*x + 4) / 5 }
integrate(fx1, lower = 0, upper = 1/2)
```

3) 확률변수의 기대값과 성질

기대값
> 확률변수 X 가 연속적일 때, 그 기대값은 X 의 값과 확률밀도함수 f (x) 를 곱한 후, 적분한 값

모집단 / 표본집단

기대값의 성질
> E(a) = a
> E(aX) = a * E(X)
> E (aX + b) = a * E(X) + b

4) 예제 : 동전던지기를 3회 시행하고 중단. 시행 중 앞면이 나오면 중단. 던진 횟수에 대한 기대값은 ?

S	(H)	(T, H)	(T, T, H) or (T, T, T)
X	1	2	3
P (X = x)	1/2	1/4	2/8

5) Variance : 분산

> 확률분포의 흩어진 정도를 나타낸다.
> 흩어진 정도의 기준은, 모평균이 된다.

분산 ** 0.5 = 표준편차

6) 분산의 성질

> Var(a) = 0

> Var (aX) = a^2 Var (X)

> Var (aX + b) = a^2 Var (X)

변동계수 : 표준편차 / 모평균

7) Covariance : 공분산

ex) 나이 - 연봉의 선형 관계 : 양의 관계 / 음의 관계 / 무관계

> 무관계 : 선형 관계가 없는 것. 하지만 비선형 관계는 존재할 수 있다.
> 양의 관계 : A 의 증가에 따라 B 가 증가. Cov(A, B) > 0
> 음의 관계 : A 의 증가에 따라 B 가 감소. Cov(A, B) < 0

8) Correlation Coefficient : 상관계수

> 공분산의 단점 : 단위 종속적이므로, 다른 공분산과의 비교가 불가능하다. 선형관계의 밀접도를 비교할 수 없다.
> 상관계수 p = Corr (X, Y) = Cov (X, Y) / { sqrt(Var (X)) * sqrt (Var (Y)) }

상관계수의 성질
> -1 <= Corr (X, Y) <= 1
> 양의 상관 / 음의 상관으로 존재한다.
> 관계 정도는 절댓값을 바탕으로 비교한다. 음수 - 양수는 방향성만을 보여줄 뿐이며, 크기 비교에서는 의미없다.

9) 두 변수의 독립성

두 확률변수 X 와 Y 가 서로 독립일 필요충분조건
f (x , y) = fx (x) fy (y)

위의 조건을 만족하는 독립 확률변수라면,
E(XY) = E(X) E (Y)
Cov(X, Y) = 0 이다.

10) 이항분포 함수

dbinom (x , size , prob)
> 확률질량함수

pbinom (q, size, prob, lower.tail = T)
> q 에 대한 binomial 함수

qbinom (a, size, prob, lower.tail = T)
> 누적확률이 a 가 되는 값 Xa 를 찾는 함수

rbinom (k, size, prob)
> 확률변수 x 를 임의로 n 개 뽑아내어 그 연산 결과를 반환하는 함수

11) 예제 : 성공확률 40%, 확률변수 X 는 15회에 대한 성공 횟수, X = 10의 확률과 X <= 2의 확률은 ? 또한 기대값과 분산은 ?

X ~ B (15, 0.4)
P (X = x) = (x C 15) 0.4 ^ x * 0.6 ^ 15 - x

```
n <- 15
p <- 0.4
dbinom ( 10, n, p )
pbinom ( 2, n, p )
Ex <- n*p
VarX <- n*p*(1-p)
```

정규분포

왜도 / 첨도 : 분포가 어느방향으로 어느정도 치우쳐져 있는지 판단할 수 있다.
왜도 = 0 : 좌우대칭을 의미한다.

직접 적분하여 확률값을 구하는 것은 매우 복잡하므로, 표준정규분포로 변환하여 구할 수 있다.

pnorm (1.5, 2, sqrt (4), lower.tail = F)
= 1 - pnorm (1.5, 2, sqrt(4))

표본분포

1) Population : 모집단

> 통계적인 관찰의 대상이 되는 집단 전체 / 모든 개체의 집단을 의미함
> 유한모집단 / 무한모집단 : 모집단의 크기가 유한하거나 무한한 경우를 의미함

2) Sample : 표본

> 모집단을 대표할 수 있도록 선택된 모집단 구성단위의 일부
> 전수조사가 시간적, 경제적 여건상 불가능한 경우. 관심 특성치가 파괴되어야 얻을 수 있는 자료의 경우
> 전수조사를 하면 오차개입이 커지므로, 표본만 구하여 조사하는 것이 합당하다

3) Prameter : 모수

4) 확률표본

> 미지의 모집단이 확률분포 F 를 따를 때, 모집단에서 추출한, 크기가 n인 확률표본 X1 ~ Xn 은
> 모집단과 동일한 확률분포 F 로부터 서로 독립적으로 추출한 n개의 확률집단을 뜻한다

5) 중심극한정리

> 비정규모집단에서의 표본평균의 분포

6) 카이제곱분포

pchisq (1.25, 3)

7) T - 분포

> 모분산을 모르는 경우에, T 분포로 추측한다.

8) F - 분포

9) 기하분포 / 초기하분포 / 균등분포

자료의 요약

1) 대표값

산술평균
> 표본평균 / 모평균을 구하여 대표값으로 정할 수 있다.
> 극단값 / 이상값에 영향을 많이 받는다

중앙값
> 자료를 순서대로 나열할 때, 가운데 값을 의미한다.
> 평균에 비해 극단값의 영향을 덜 받는다.

최빈값
> 가장 빈도가 많은 값을 의미한다.
> 빈도에 의한 것이므로, 여러개의 최빈값이 존재하거나, 아예 존재하지 않을 수 있다

절사평균
> 상한 / 하한 n% 만큼의 값을 제거하고, 나머지 값을 이용해 평균값을 구하는 것
> 평균의 장점과 중앙값의 장점을 모두 갖는 대표값이다

2) R의 대표값 연산 함수

maen(x, trim = 0, na.rm = T)
> 산술평균 함수

median(x, na.rm = T)
> 중앙값 함수

Mode <- function (x) { ux <- unique (x) ; ux [which . max (tabulate (match (x, ux)))]
> 최빈값 함수 : 존재하지 않으므로 직접 만들어 사용한다.

3) 산포도

> 데이터가 대표값을 기준으로 얼마나 흩어져있는가를 나타내는 수치

범위
> 최대값과 최소값의 차이. 데이터가 퍼져 있는 정도를 나타내는 가장 간단한 방법
> 이상치가 하나만 존재하여도 너무 넓은 범위가 나타날 수 있다

사분위수 편차
> 데이터를 크기순서로 나열한 다음, 개수로 4등분했을 때 1사분위수와 3사분위수의 차이를 2로 나눈 것
> 즉, 4분위로 나눈 상태에서, 하위 25%와 상위 25% 값을 모두 버렸을 때, 최대 - 최소 값을 의미한다.

분산
> 산포도의 척도로써 가장 널리 사용됨. 데이터가 퍼져있는 정도의 기준으로 평균을 사용한다.
> 표본분산 / 모분산 등이 존재한다.
> 분산 / 표준편차가 0이라면, 모든 데이터가 동일하다는 의미이다

변동계수
> 서로 다른 단위수를 갖는 자료들의 산포를 비교할 때 사용된다.

4) 위치척도

최대값과 최소값
> 크기 순서로 늘어놓은 자료에서, 가장 큰 값과 가장 작은 값

사분위수
> 크기 순서로 늘어놓은 자료를 4등분하는 수 : 1, 2, 3, 4분위수가 존재

5) R의 산포도 연산 함수

var(x, na.rm = F)
> 분산 함수

sd(x, na.rm = F)
> 표준편차 함수

cov(x, y = NULL)
> 공분산 함수
> cov 대신에 var 를 이용해도 동일한 결과를 얻을 수 있다.

range(x)
> 최대값 - 최소값의 범위 함수

IQR(x, na.rm = F)
> 사분위범위 함수

quantile(x, probs = seq(0, 1, 0.25), na.rm = F, names = T)
> 분위수 함수. probs 를 이용해서 출력 범위를 조정한다

summary(x)
> 핵심적인 자료 요약값을 보여줌