SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 도수분포표(Frequency table)

- 범주형 자료 또는 범주화된 자료를 정리
- 각 범주에 몇 개의 관측개체가 있는지를 정리한 표
  - 도수
  - 상대도수

⇒ 표본을 계속 뽑으면

⇒ **해당범주가 모집단에서 차지하는 비율(확률)**로 수렴

↳ 통계학에서의 관심사

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 다항분포(Multinomial Distribution)

- 각 시행에서 발생 가능한 결과는 k가지
- 각 시행에서 i 번째 결과의 확률은 p_i로 고정, $\sum_{i=1}^{k} p_i = 1$
- 각 시행은 독립적으로 수행

- $(X_1, X_2, \ldots, X_k)$: n번 시행했을 때, 각 결과의 횟수

시행	결과 1	결과 2	⋯	결과 k	합
1	X_{11}	X_{12}	⋯	X_{1k}	1
2	X_{21}	X_{22}	⋯	X_{2k}	1
⋮	⋮	⋮	⋱	⋮	⋮
n	X_{n1}	X_{n2}	⋯	X_{nk}	1
합	X_1	X_2	⋯	X_k	n

  - X_{ij}: i번째 시행에서 결과 j가 나오면 1 아니면 0
    - $X_{ij} = 1$이면 $X_{il} = 0,\ l \neq j$
  - $i_1 \neq i_2$인 경우 $X_{i_1 j_1}$와 $X_{i_2 j_2}$는 서로 독립

- 이항분포: $X_1 \sim B(n, p_1)$

$$f(x_1) = \frac{n!}{x_1!(n-x_1)!} p_1^{x_1} (1-p_1)^{n-x_1}$$

  - $X_2 = n - X_1,\ p_2 = 1 - p_1$

$$f(x_1, x_2) = \frac{n!}{x_1! x_2!} p_1^{x_1} p_2^{x_2}, \qquad x_1 + x_2 = n$$

- 다항분포의 확률질량함수

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

  - $\sum_{i=1}^{k} x_i = n,\ \sum_{i=1}^{k} p_i = 1$

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

◉멘델의 유전법칙

◦ 독립의 법칙: 완두의 껍질 모양(R,r), 색깔(Y,y)

· RRYY, rryy인 완두 교배 1대를 자기수분시킨 2대의 발현비율

RY:Ry:rY:ry = 9:3:3:1

◦ 독립적으로 n개의 2대를 얻었을 때, (RY,Ry,rY,ry)에 속한 완두의 수를 (X_1, X_2, X_3, X_4)라고 하면

$$f(x_1,x_2,x_3,x_4) = \frac{n!}{x_1!x_2!x_3!x_4}\left(\frac{9}{16}\right)^{x_1}\left(\frac{3}{16}\right)^{x_2}\left(\frac{3}{16}\right)^{x_3}\left(\frac{1}{16}\right)^{x_4}$$

- 특정 결과에만 관심이 있는 경우,
  - 예】 i-번째 결과(R_i)에만 관심 ⇒ 나머지 결과를 묶음(R_i^c)

$$X_i \sim B(n, p_i)$$

- $E(X_i) = np_i$
- $Var(X_i) = np_i(1-p_i)$

  - 예】i-번째 또는 j-번째 결과($R_i \cup R_j$) 관심

$$Y = X_i + X_j \sim B(n, p_i + p_j)$$

- $E(Y) = E(X_i + X_j) = n(p_i + p_j)$
- $Var(Y) = Var(X_1 + X_2) = n(p_i + p_j)(1-(p_i + p_j))$

- X_i와 X_j와의 관계
  - $Cov(X_{11}+X_{21}, X_{12}+X_{22})$

    $= Cov(X_{11}, X_{12}) + Cov(X_{11}, X_{22}) + Cov(X_{21}, X_{12}) + Cov(X_{21}, X_{22})$
  - $Cov(X_1, X_2) = \sum_{i=1}^{n} Cov(X_{i1}, X_{i2})$
  - $Cov(X_{i1}, X_{i2}) = E(X_{i1}X_{i2}) - E(X_{i1})E(X_{i2})$
    - $E(X_{ij}) = p_j,\ E(X_{i1}X_{i2}) = 0$

      $\Rightarrow\ Cov(X_{i1}, X_{i2}) = -p_1 p_2$
  - $Cov(X_i, X_j) = -n p_i p_j$

- $$Cor(X_i, X_j) = \frac{-np_ip_j}{\sqrt{np_i(1-p_i)}\sqrt{np_i(1-p_i)}}$$
$$= -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}$$

  · $p_i/(1-p_i)$: 오즈(odd)

- $$Var(X_i + X_j) = Var(X_i) + Var(X_j) + 2Cov(X_i, X_j)$$
$$= np_i(1-p_i) + np_j(1-p_j) - 2np_ip_j = n\left(p_i + p_j - (p_i + p_j)^2\right)$$
$$= n(p_i + p_j)\left(1 - (p_i + p_j)\right)$$

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

◉멘델의 유전법칙

◦ 모양(R, r)에만 관심이 있는 경우, R:r = 12:4 = 3:1

· R의 개수: $Y = X_1 + X_2 \sim B(n, 0.75)$

◦ 100개의 완두에 대해 우성인자만 있는 경우와 열성인자만 있는 완두 수의 상관계수는?

- $p_1 = 9/16,\ p_4 = 1/16$
- $Cov(X_1, X_4) = -160 \times (9/16) \times (1/16) = -5.625$
- $Cor(X_1, X_4) = -\sqrt{\dfrac{(9/16)(1/16)}{(7/16)(15/16)}} = -0.2928$

- **cf.** $Cor(X_1, X_2) = -\sqrt{\frac{(9/16)(3/16)}{(7/16)(13/16)}} = -0.5447$

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

- 요약
  - 각 시행에서 발생 가능한 결과는 k가지
  - 각 시행에서 i번째 결과의 확률은 p_i로 고정,
  - 각 시행은 독립적으로 수행
  - n번 시행했을 때 각 결과의 횟수 분포
  - 특정 결과의 횟수의 분포 ⇒ 이항분포
  - $$Cov(X_i, X_j) = -np_ip_j$$
  - $$Cor(X_i, X_j) = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}$$