Lecture 09 - 확률과 분포 이해하기

개요

- 표본 공간, 사건, 확률 이해하기
- 확률 질량 함수와 확률 밀도 함수 이해하기
- 확률 분포를 이용해서 확률 계산하기

표본 공간(sample space)

- 표본 공간이란 어떤 실험(experiment)이나 시행(trial)의 모든 가능한 결과를 모아 놓은 집합이다.

표본 공간의 예시

동전을 한 번 던지는 시행

앞면, 뒷면 2개의 결과가 가능
{head, tail}

동전을 두 번 던지는 시행

{(head,head), (head,tail), (tail,head), (tail,tail)}

주사위를 한 번 던지는 시행

{⚀, ⚁, ⚂, ⚃, ⚄, ⚅}

표본 공간이 중요한 이유

- 확률은 불확실성(uncertainty)을 갖는 실험이나 시행의 개별 결과가 나타날 가능성을 의미
- 가능한 결과를 모두 정의할 수 있어야 확률을 제대로 정의할 수 있다.

사건(event)

- 사건이란 실험이나 시행의 결과를 모아 놓은 집합이다.
- 표본 공간의 부분 집합

사건의 예시

동전을 한 번 던지는 시행에서 앞면이 나오는 사건

{head}

동전을 두 번 던져서 앞 면이 두 번 나오는 사건

{(head, head)}

주사위를 한 번 던져서 6이 나오는 사건

{⚅}

사건이 중요한 이유

- 실제로 관심이 있는 사건은 표본 공간의 일부
- 관심 있는 사건의 확률을 구하기 위해서 필요

확률(probability)

- 확률이란 어떤 사건이 일어날 수 있는 가능성에 대한 척도다.
- 0과 1사이의 값을 가지며, 값이 큰 사건은 값이 작은 사건보다 일어날 가능성이 높다.

확률 계산

- 표본 공간에서 사건이 차지하는 비율이 곧 확률

확률 계산 예시

동전을 한 번 던져서 앞면이 나올 확률

- 동전을 한 번 던지는 시행의 표본 공간 = Ω = {head, tail}
- 동전을 한 번 던졌을 때 앞면이 나오는 사건 = A = {head}
- 동전을 한 번 던졌을 때 앞면이 나올 확률 = $P(A) = \frac{|A|}{|\Omega|} = \frac{1}{2}$

주사위를 한 번 던져서 6이 나올 확률

- 주사위를 한 번 던지는 시행의 표본 공간 = Ω = {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}
- 주사위를 한 번 던졌을 때 6이 나오는 사건 = A = {⚅}
- 주사위를 한 번 던졌을 때 6이 나올 확률 = $P(A) = \frac{|A|}{|\Omega|} = \frac{1}{6}$

벤다이어그램을 이용한 확률 계산

- 표본 공간과 사건은 집합이므로 벤다이어그램(Venn diagram)을 이용한 확률 계산이 가능
- 아래 그림에서 표본 공간이 B라면, 사건은 A가 된다.(사건은 표본 공간의 부분 집합)
- 두 영역의 넓이 비율이 A:B = 1:0.4라면 사건 B가 일어날 확률은 0.4

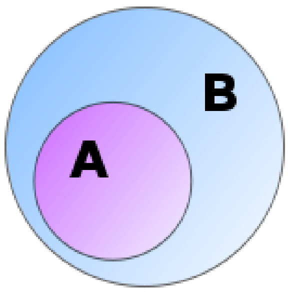

그림 출처: [wikipedia]

확률 변수(random variable)

- 표본 공간의 원소(=근원 사건)를 실수와 연결하는 일종의 함수

확률 변수의 예시

동전을 한 번 던지는 시행: $\Omega \rightarrow X$

- $X = 1 \ if \ \omega = head$
- $X = 0 \ if \ \omega = tail$

- 특정 사건에 대한 확률: $P(X = x)$
- 동전을 한 번 던졌을 때, 앞면이 나올 확률: $P(X = 1)$
- 동전을 한 번 던졌을 때, 뒷면이 나올 확률: $P(X = 0)$

주사위를 한 번 던지는 시행: $\Omega \rightarrow Y$

- $Y = 1 \; if \; \omega =$ ⚀
- $Y = 2 \; if \; \omega =$ ⚁
- $Y = 3 \; if \; \omega =$ ⚂
- $Y = 4 \; if \; \omega =$ ⚃
- $Y = 5 \; if \; \omega =$ ⚄
- $Y = 6 \; if \; \omega =$ ⚅
- 특정 사건에 대한 확률: $P(Y = y)$
- 주사위를 한 번 던졌을 때, 6이 나올 확률: $P(Y = 6)$

확률 변수가 중요한 이유

- 현실 세계의 사건을 수학 세계의 숫자로 치환
- 수학의 좋은 방법들을 확률에 적용할 수 있다.

확률 분포(probability distribution)

- 확률 분포란 확률 변수 값(사건)에 확률을 나눠준 것.
  - 모든 가능한 사건들의 확률 총합은 1.
  - 1이라는 가능성을 모든 사건들에게 적절히 나누어 주는(distribute) 것.
  - 함수로 표현하면 편하다.

확률 질량 함수(probability mass function)

- 이산 확률 변수(discrete random variable)
- mass = probability

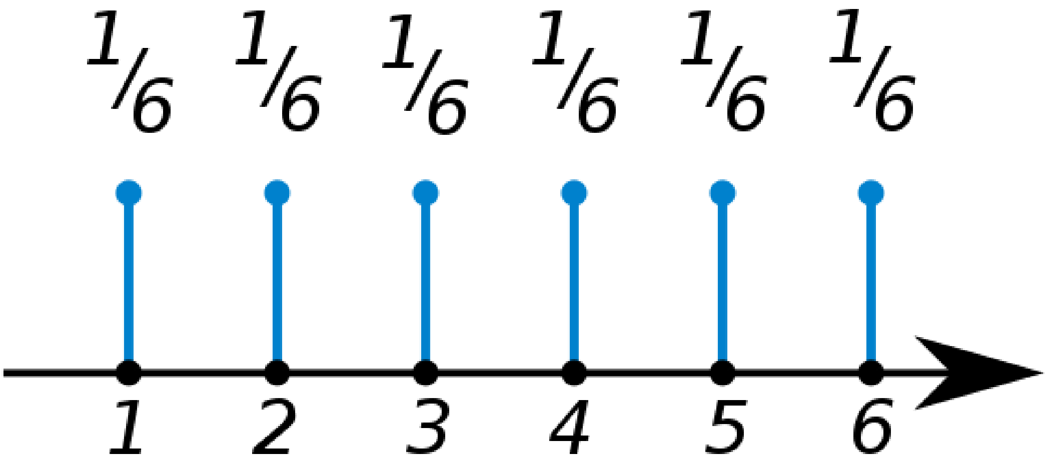

그림 출처: [wikipedia]

확률 밀도 함수(probability density function)

- 연속 확률 변수(continous random variable)
- density = mass / volume
  mass = density * volume => concept of integral

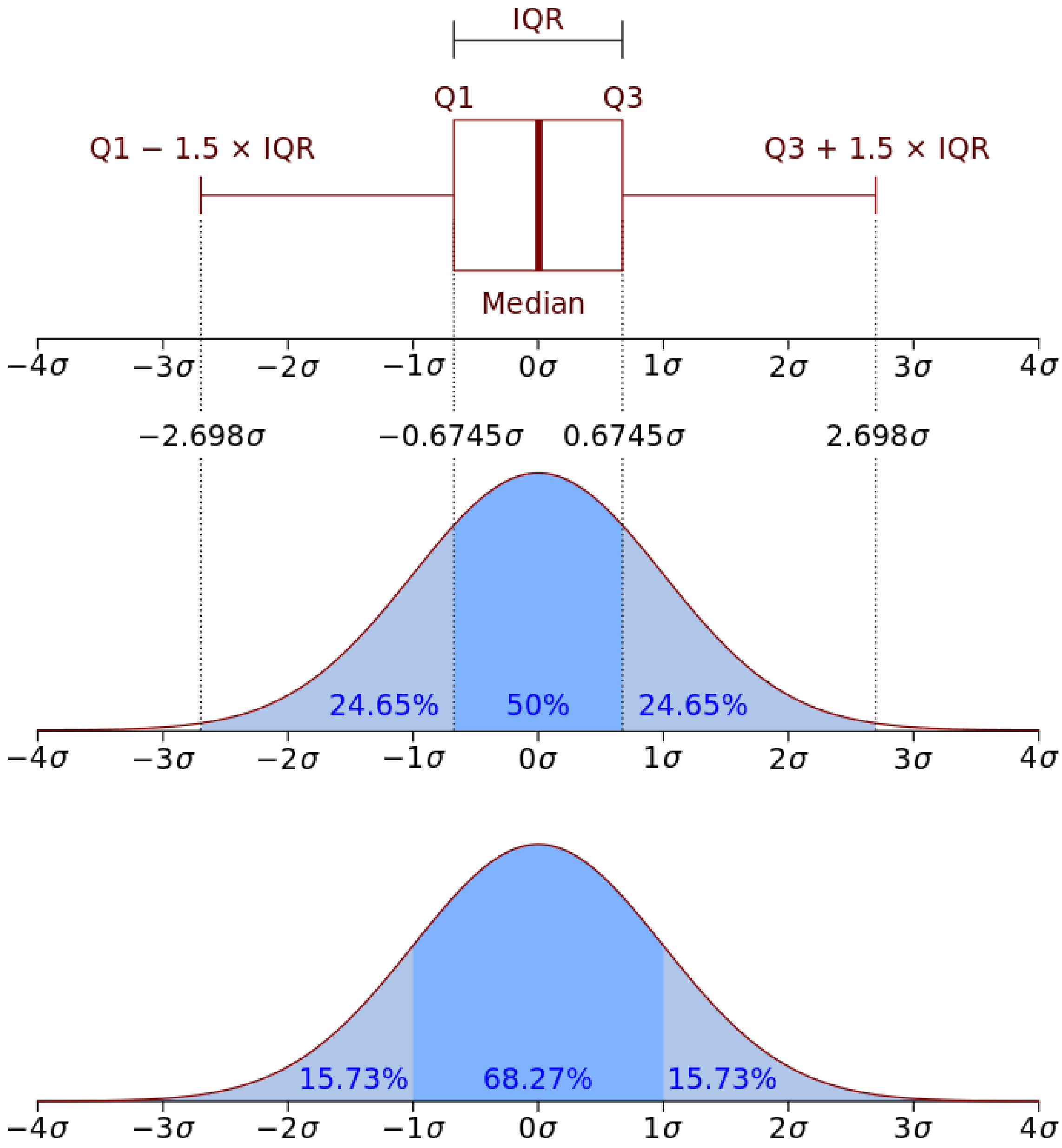

그림 출처: [wikipedia]

참고: [List of probability distribution]

확률 분포 종류

- 설명하고자 원하는 시행에 따라서 아주 많은 확률 분포가 존재한다.
- 가장 대표적으로 '이항 분포'와 '정규 분포'를 많이 사용.
- 참고: 여러가지 확률 분포 https://en.wikipedia.org/wiki/Probability_distribution#Common_probability_distributions

R에서의 확률 분포

- 기본적으로 10 여개의 확률 분포에 대해서 계산
    - https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Distributions.html
- 각 확률 분포에 대해서 아래 4개를 계산하는 함수가 존재
    - density/mass function: **d**xxx
    - cumulative distribution function: **p**xxx
    - quantile function: **q**xxx
    - random variate generation: **r**xxx
- 예를 들어,
    - 이항 분포(binom): **d**binom, **p**binom, **q**binom, **r**binom
    - 정규 분포(norm): **d**norm, **p**norm, **q**norm, **r**norm

이항 분포(binomial distribution)

- bi + nomial(cf. multi + nomial)
- 결과가 단 2개(성공, 실패) 뿐이고, 성공할 확률이 p인 시행을 n번 시도했을 때, x번 성공할 확률
- $X \sim B(n, p)$

$$f(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

확률 계산 예시

당첨 확률이 1%인 즉석 복권을 100번 사서 0번 당첨될 확률은?

```
dbinom(0, 100, .01)
```

당첨 확률이 1%인 즉석 복권을 100번 사서 1번 당첨될 확률은?

```
dbinom(1, 100, .01)
```

당첨 확률이 1%인 즉석 복권을 100번 사서 2번 당첨될 확률은?

```
dbinom(2, 100, .01)
```

당첨 확률이 1%인 즉석 복권을 100번 사서 최소한 2번 이하로 당첨될 확률은?

```
pbinom(2, 100, .01)
```

정규 분포(normal distribution)

- 자연 과학이나 사회 과학에서 실측 값에 대한 분포로 자주 사용
- $X \sim N(\mu, \sigma)$

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

확률 계산 예시

한국 성인 남성 키 평균 174, 표준 편차 6
한국 성인 여성 키 평균 161, 표준 편차 5

키 170인 성인 남성은 하위 몇 %인가?

```
pnorm(170, 174, 6)
```

키 180인 성인 남성은 하위 몇 %인가?

```
pnorm(180, 174, 6)
```

키 180인 성인 남성은 상위 몇 %인가?

```
1 - pnorm(180, 174, 6)
pnorm(180, 174, 6, lower.tail=F)
```

성인 여성의 키가 175일 확률은?

```
dnorm(175, 161, 5)  # wrong answer
```

성인 여성의 키가 175이상일 확률은?

```
1 - pnorm(175, 161, 5)
```

성인 여성의 키가 160이상 170미만일 확률은?

```
pnorm(170, 165, 5) - pnorm(160, 165, 5)
```

성인 남성 상위 5%의 키는?

```
qnorm(.95, 174, 6)
```

경험 확률 분포(empirical probability distribution)

- 특정 사건이 일어난 확률을 ‘사건 관찰 수 / 총 시행 수’로 구하는 경우, 이를 경험 확률이라고 한다.
- = 상대 도수(relative frequency)
- 0에 아주 가까운 확률을 갖는 사건에 대해서 경험 확률을 구하기 어렵다. (아주 많은 시행 수가 필요)
- 적당한 확률 분포를 가정할 수 없을 때 사용 가능

계산이 어려운 확률 문제

확률 시뮬레이션 예시

365일 동안 매일 의사결정을 하는 회사가 있다.

- .51 회사는 옳은 의사 결정을 할 확률이 51%
- .49 회사는 옳은 의사 결정을 할 확률이 49%
- 옳은 의사 결정을 내렸을 때의 이익은 0~2까지 무작위
- 나쁜 의사 결정을 내렸을 때의 손해는 -1 고정
- 1년 후에 각 회사의 누적 이익은 얼마일까?

구현 코드: simple_simulation.R