확률과 확률변수

□ 학습 목표

- 확률의 개념과 계산 원리를 이해한다.
- 확률변수와 확률분포(이산형/연속형)를 구분할 수 있다.
- 베르누이, 이항, 포아송, 정규, 지수분포의 의미와 차이를 설명할 수 있다.
- Python 실습으로 간단한 분포 시뮬레이션을 경험한다.

1. 확률의 개념

1) 확률이란?

어떤 사건이 일어날 가능성을 0과 1 사이의 수로 표현한 것

0 = 절대 일어나지 않음, 1 = 반드시 일어남

2) 표본공간과 사건

- 표본공간(S): 가능한 모든 결과들의 집합
- 사건(A): 우리가 관심 있는 특정 결과

예시

- 주사위 던지기: S = {1,2,3,4,5,6}
- A = {짝수가 나옴} = {2,4,6}

2. 확률변수 (Random Variable)

1) 정의

사건의 결과를 숫자로 바꾼 것

X = “동전 던져 앞면이면 1, 뒷면이면 0”

2) 종류

- 이산형: 값이 딱딱 떨어지는 경우 (0,1,2,…) → 확률질량함수(PMF)
- 연속형: 값이 연속적으로 나타나는 경우 (시간, 길이) → 확률밀도함수(PDF)

3. 주요 확률분포

1) 베르누이 분포

- 한 번의 시행 (성공/실패)
- P(X=1)=p, P(X=0)=1-p

📌 예시 문제 2

불량률 문제

- 어떤 공장에서 만든 부품이 불량일 확률 = $p = 0.05$
- 하나의 부품을 뽑았을 때,
  1. 불량일 확률?
  2. 정상일 확률?

풀이

- $P(X = 1) = 0.05$ (불량)
- $P(X = 0) = 0.95$ (정상)

✅ 답: 불량 확률 5%, 정상 확률 95%

📌 예시 문제 3 (확장)

- 시험 문제를 맞출 확률 = $p = 0.8$
- 문제 1개를 찍었을 때, 정답일 확률과 오답일 확률은?

풀이

- $P(X = 1) = 0.8$
- $P(X = 0) = 0.2$

📊 정리

- 베르누이 분포는 **항상 2가지 결과(0,1)**
- 확률은 **p와 1-p** 두 개
- 평균 = $E[X] = p$, 분산 = $Var(X) = p(1-p)$

2) 이항 분포

- n번 반복한 베르누이 시행 → 성공 횟수
- 평균 = np, 분산 = np(1-p)

예시: 동전 10번 던져 앞면 개수 3개일 확률

📌 **문제 정리**

- 동전은 앞면 확률 $p = 0.5$
- 시행 횟수 $n = 10$
- 성공(앞면) 횟수 $X \sim Binomial(n = 10, p = 0.5)$
- 구할 확률:

$$P(X = 3) = \binom{10}{3}(0.5)^3(0.5)^7$$

📌 **계산**

- $\binom{10}{3} = \frac{10!}{3!(10-3)!} = 120$
- 따라서:

$$P(X = 3) = 120 \times (0.5)^{10}$$
$$= 120 \times \frac{1}{1024} \approx 0.117$$

3) 포아송 분포

- 일정한 시간/공간 내 사건 발생 횟수
- 모수 λ = 평균 발생 횟수

예시: 콜센터에 1시간 평균 6콜이 올 때, 30분 동안 전화가 없을 확률?

📌 **문제 정리**

- 1시간 평균 = λ = 6콜
- 30분 = 0.5시간 → 평균 발생 횟수 = λt = 6 × 0.5 = 3콜
- X ~ Poisson(3)
- 구할 확률:

$$P(X = 0) = \frac{e^{-3}3^0}{0!}$$

📌 **계산**

$$P(X = 0) = e^{-3} = 0.0498 \approx 4.98\%$$

■ 중심극한정리

📖 중심극한정리 (Central Limit Theorem)

중심극한정리란, 동일한 확률분포를 따르는 독립적인 확률변수들의 표본을 충분히 많이 추출할 경우, 원래의 분포가 어떠하든지 간에 그 표본평균의 분포가 점점 정규분포에 가까워진다는 정리를 말한다.

즉, 모집단의 분포가 정규분포가 아니더라도, 표본의 크기 n이 충분히 크면 표본평균 $\bar{X}$는 평균이 μ, 분산이 σ^2/n인 정규분포를 근사적으로 따른다.

수식으로 표현하면 다음과 같다.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1), \quad (n \to \infty)$$

여기서 μ는 모집단의 평균, σ^2는 모집단의 분산을 의미한다.

4) 지수 분포

- 포아송 과정에서 사건 간 시간 간격을 나타내는 확률분포.

📌 확률밀도함수 (PDF)

$$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$

- $\lambda > 0$: 사건 발생률 (단위 시간당 평균 발생 횟수)
- t: 사건이 발생하기까지 걸린 시간
- 성질:
  - $f(t) \geq 0$
  - $\int_0^\infty f(t)\,dt = 1$ (전체 확률 = 1)

📌 누적분포함수 (CDF)

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t}, \quad t \geq 0$$

- "t 시간 이내에 사건이 발생할 확률"

📌 기대값과 분산

- 기대값(평균):

$$E[T] = \frac{1}{\lambda}$$

- 분산:

$$Var(T) = \frac{1}{\lambda^2}$$

👉 평균 간격이 1/λ이라는 의미 = 사건이 자주 일어나면 (λ 크면) 평균 간격이 짧고,
드물게 일어나면 (λ 작으면) 평균 간격이 길어진다는 뜻

📊 직관적인 예시

- 콜센터에 **평균 1시간에 6콜** → λ=6
  - 평균 도착 간격 = 1/6 시간 = 10분
  - PDF: $f(t) = 6e^{-6t}$
  - "다음 전화가 5분 안에 올 확률"

$$P(T \leq 0.083) = 1 - e^{-6 \cdot 0.083} \approx 0.393$$

문제 : 어떤 기계의 평균 고장 간격 시간이 200시간이라고 한다.
즉, 고장까지 걸리는 시간 T는 지수분포 Exp(λ)를 따른다.

1. λ의 값을 구하시오.
2. 이 기계가 100시간 이상 정상적으로 작동할 확률은 얼마인가?
3. 이 기계가 300시간 이내에 고장날 확률은 얼마인가?

1. **평균=1/λ**

$$E[T] = \frac{1}{\lambda} = 200 \quad \Rightarrow \quad \lambda = \frac{1}{200} = 0.005$$

2. **100시간 이상 정상 작동**

$$P(T > 100) = e^{-\lambda \cdot 100} = e^{-0.005 \times 100} = e^{-0.5} \approx 0.6065$$

3. **300시간 이내 고장**

$$P(T \leq 300) = 1 - e^{-\lambda \cdot 300} = 1 - e^{-0.005 \times 300} = 1 - e^{-1.5} \approx 0.7769$$

5) 정규 분포

- 평균을 중심으로 대칭, 종 모양 곡선
- 많은 자연현상과 데이터가 근사적으로 따름

예시: 어떤 반의 수학 시험 점수는 평균이 70점, 표준편차가 10점인 정규분포를 따른다.
확률변수 X~N(70, 10^2)

1) 학생이 80점 이상을 받을 확률은?
2) 학생이 60점 이상 80점 이하를 받을 확률은?
3) 상위 5%에 드는 학생의 점수 기준(컷라인)은 몇 점 이상인가?

1. **표준화(Z-score)**

$$Z = \frac{X - \mu}{\sigma}$$

---

(1) 80점 이상

$$P(X \geq 80) = P\left(Z \geq \frac{80 - 70}{10}\right) = P(Z \geq 1.0)$$

표준정규분포표에서 $P(Z \geq 1.0) \approx 0.1587$

👉 답: 약 **15.9%**

---

(2) 60 ≤ X ≤ 80

$$P(60 \leq X \leq 80) = P\left(-1 \leq Z \leq 1\right)$$

$$= P(Z \leq 1) - P(Z \leq -1) = 0.8413 - 0.1587 = 0.6826$$

👉 답: 약 **68.3%**

---

(3) 상위 5%

상위 5% → 누적확률 0.95에 해당하는 Z값 ≈ 1.645

$$X = \mu + Z\sigma = 70 + 1.645 \times 10 = 86.45$$

👉 답: 약 **87점 이상**

6. 문제:

1) 불량률이 5%인 부품에서 20개를 뽑았을 때 불량이 2개 이하일 확률?

📌 문제 정리

- 각 부품이 불량일 확률 = p = 0.05
- 뽑는 개수 = n = 20
- 불량품의 개수 = X ~ Binomial(n=20, p=0.05)
- 구할 확률:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

📌 계산

이항분포 확률질량함수(PMF):

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- $P(X = 0) = \binom{20}{0}(0.05)^0(0.95)^{20} \approx 0.358$
- $P(X = 1) = \binom{20}{1}(0.05)^1(0.95)^{19} \approx 0.377$
- $P(X = 2) = \binom{20}{2}(0.05)^2(0.95)^{18} \approx 0.189$

📊 합계

$$P(X \leq 2) \approx 0.358 + 0.377 + 0.189 = 0.924$$

추정과 가설검증

1. 추정 (Estimation)

1) 개념

- 모집단의 특성(모수: 평균 μ, 비율 p 등)을 직접 알 수 없을 때,
- 표본 데이터를 이용해 그 값을 추정하는 것.

2) 종류

- 점추정 (Point estimation)

  · 하나의 값으로 추정
  예: 표본평균 = 68점 → 모집단 평균 μ 추정

- 구간추정 (Interval estimation)
  · “얼마나 오차가 있을지”까지 반영 → 신뢰구간
  예: 95% 신뢰수준에서 $\mu \in [65, 71]$

✓ 점추정은 단순, 구간추정은 신뢰감 부여

2. 가설검증 (Hypothesis Testing)

1) 개념

· 모집단에 대해 세운 주장(가설)이 옳은지 틀린지 표본 데이터로 검증하는 과정

2) 절차

- 가설 설정

  · 귀무가설(H_0): 변화 없음, 차이 없음 (기본 가설)
  · 대립가설(H_1): 변화 있음, 차이 있음 (검증 대상)

- 유의수준(α) 결정

  · 보통 0.05 (5%) → “우연히 이런 결과가 나올 확률이 5% 이하라면 H_0 기각”

- 검정통계량 계산

  · 표본 데이터를 이용해 Z, t, χ^2, F 등 산출

- 기각역과 비교

  · 검정통계량이 기각역에 들어가면 H_0 기각, 아니면 채택

3) 예시

- 신약 실험: “신약 효과가 없다(H_0)” vs “신약 효과가 있다(H_1)”

- 표본 실험 결과가 우연히 발생할 확률이 매우 작으면 → 신약 효과 있다고 결론

3. 추정 vs 가설검증 비교

구분	추정	가설검증
목적	모수의 값을 알아내는 것	어떤 주장이 맞는지 검증
방법	점추정, 구간추정	귀무가설, 대립가설 설정 후 검정
결과	“평균은 65~71 사이일 것이다”	“평균이 70이라는 가설은 기각된다”

예시 문제

1. 추정 (Estimation)

문제 1. 점추정과 구간추정

- 어떤 학급에서 무작위로 학생 25명을 뽑아 수학 점수를 조사한 결과, 표본평균이 68점이고 표본표준편차가 10점이었다.

1) 모집단 평균 점수에 대한 점추정값은 얼마인가?
2) 모집단 평균에 대한 95% 신뢰구간을 구하시오.

풀이

1. 점추정: $\hat{\mu} = \bar{X} = 68$
2. 구간추정:

$$\bar{X} \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

- $\bar{X} = 68, s = 10, n = 25, Z_{0.025} = 1.96$
- 오차한계 = $1.96 \times \frac{10}{\sqrt{25}} = 1.96 \times 2 = 3.92$
- 신뢰구간 = [64.08, 71.92]

👉 95% 신뢰수준에서 모집단 평균은 **64.1점 ~ 71.9점**으로 추정된다.

2. 가설검정 (Hypothesis Testing)

문제 1. 평균에 대한 가설검정

- 한 제약회사는 신약이 혈압을 낮추는 효과가 있다고 주장한다.
기존 약물의 평균 혈압 감소량은 5mmHg인데, 신약을 투여한 36명의 표본에서 평균 혈압 감소량이 6mmHg, 표준편차가 3mmHg로 나타났다.
유의수준 5%에서 신약이 효과가 있다고 말할 수 있는가?

풀이

1. 가설 설정
- H_0: μ = 5 (효과 없음)
- H_1: μ > 5 (효과 있음)
2. 검정통계량

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{6 - 5}{3/\sqrt{36}} = \frac{1}{0.5} = 2.0$$

3. 기각역
- 유의수준 0.05에서 단측검정
- $Z_{0.05} = 1.645$
4. 결론
- 계산된 Z=2.0 > 1.645 → 귀무가설 기각
- 신약이 효과 있다는 증거가 충분하다.

가설검증 (Hypothesis Testing)

1. 가설검증이란?

- 모집단에 대한 어떤 주장이 맞는지, 표본 데이터를 이용하여 통계적으로 검토하는 절차.
- 단순히 평균 계산이 아니라 “주장이 사실인지” 판단하는 방법.

2. 가설검증의 절차

(1) 가설 설정

- 귀무가설(H_0): 변화 없음, 차이 없음 (현상 유지 가설)
- 대립가설(H_1): 변화 있음, 차이 있음 (검증하고 싶은 주장)

□ 예시:

- 신약 효과 검증 →

· H_0: 신약 효과 없음 ($\mu = 5$)
· H_1: 신약 효과 있음 ($\mu > 5$)

(2) 유의수준(α) 결정

- 보통 0.05 (5%) 사용

“우연히 이런 데이터가 나올 확률이 5% 이하라면 H_0를 기각하자”는 기준

(3) 검정통계량 계산

- 표본 데이터를 이용해 **검정통계량(Z, t, χ^2, F 등)**을 계산

검정통계량 = “데이터가 H_0에서 얼마나 벗어났는지” 측정

(4) 기각역과 비교

- 검정통계량이 기각역(critical region)에 들어가면 → H_0 기각
- 들어가지 않으면 → H_0 채택(기각 불가)

(5) 결론 해석

- H_0를 기각하면: “데이터는 대립가설을 지지한다”
- H_0를 기각하지 못하면: “데이터로는 충분한 증거가 없다”

3. 오류의 가능성

- 제1종 오류 (α): 사실은 H_0가 참인데 기각 (잘못된 긍정)
- 제2종 오류 (β): 사실은 H_1가 참인데 H_0를 기각하지 못함 (잘못된 부정)

✓ 통계에서 “유의수준 5%”라는 말은 “제1종 오류를 5%로 제한한다”는 의미.

제1종 오류와 제2종 오류

1. 정의

□ 제1종 오류 (Type I Error)

- 실제로는 귀무가설(H_0)이 참인데, 표본 결과 때문에 잘못 기각하는 오류

  “원래 맞는 가설을 틀렸다고 판단하는 것”

- α (유의수준) = 제1종 오류를 허용하는 확률

예시:
- 신약은 사실 효과 없음 (H_0 참)
- 실험 결과 우연히 효과가 있는 것처럼 나와 → “효과 있다” 잘못 결론

□ 제2종 오류 (Type II Error)

- 실제로는 대립가설(H_1)이 참인데, 표본 결과 때문에 귀무가설을 기각하지 못하는 오류
  “틀린 가설을 그냥 두는 것”

- β = 제2종 오류 확률

예시:
- 신약은 사실 효과 있음 (H_1 참)
- 실험 표본이 작아서 차이가 잘 안 보임 → “효과 없다” 결론

2. 비유 (재판 예시)

- H_0 (귀무가설) = 피고인은 무죄다
- H_1 (대립가설) = 피고인은 유죄다

실제	판결	결과
무죄	유죄 판결	제1종 오류 (무죄인데 유죄 판결)
유죄	무죄 판결	제2종 오류 (유죄인데 무죄 판결)

· 제1종 오류는 “무죄인 사람을 잡아넣는 것”
· 제2종 오류는 “유죄인 사람을 풀어주는 것”

3. 관계와 트레이드오프

- 유의수준 α를 낮추면(=제1종 오류 줄임) → 제2종 오류는 커질 수 있음
- 표본 수 n을 크게 하면 → 두 오류 모두 줄일 수 있음

4. 요약

- 제1종 오류 (α): 잘못된 긍정 (False Positive)
- 제2종 오류 (β): 잘못된 부정 (False Negative)
- 통계에서는 보통 α=0.05 (5%)로 정해 놓고 분석을 진행