확률 분포: 이항 분포, 포아송 분포, 정규 분포

이항 분포 (Binomial Distribution)

- **이산 확률 분포**의 하나로, 서로 독립적인 베르누이 시행(결과가 성공/실패 두 가지뿐인 시행)을 n번 반복했을 때, 성공이 k번 나타날 확률을 나타냅니다.
    - **조건**: 각 시행은 독립적, 각 시행의 결과는 성공/실패 두 가지, 각 시행의 성공 확률(p)은 동일.
    - **주요 파라미터**: n (시행 횟수), p (성공 확률).
    - **예시**: 동전을 10번 던져서 앞면이 3번 나올 확률, 100개의 제품 중 불량품이 5개 나올 확률.

포아송 분포 (Poisson Distribution)

- **이산 확률 분포**의 하나로, 단위 시간 또는 단위 공간 내에서 어떤 사건이 평균적으로 λ번 발생할 때, 이 사건이 k번 발생할 확률을 나타냅니다.
    - **조건**: 각 사건은 독립적으로 발생, 단위 시간/공간 내에서 발생하는 사건의 평균 횟수는 일정.
    - **주요 파라미터**: $λ$ (람다, 단위 시간/공간당 평균 발생 횟수).
    - **예시**: 한 시간 동안 은행에 방문하는 고객 수, 1제곱미터의 천에서 발견되는 흠집의 수.

정규 분포 (Normal Distribution / Gaussian Distribution)

- **연속 확률 분포**의 하나로, 자연계와 사회 현상에서 가장 흔하게 관찰되는 종 모양(bell-shaped)의 대칭적인 분포입니다. 평균($μ$)을 중심으로 데이터가 밀집되어 있고, 평균에서 멀어질수록 데이터의 빈도가 줄어듭니다.
    - **특징**: 평균($μ$)과 중앙값(median), 최빈값(mode)이 모두 같습니다. 분포의 형태는 평균($μ$)과 표준편차($σ$)에 의해 결정됩니다.
    - **중요성**: 많은 통계적 기법(t-test, ANOVA, 선형 회귀 등)이 데이터가 정규분포를 따른다는 가정을 기반으로 합니다. 중심 극한 정리에 따라, 표본의 크기가 충분히 크면 표본 평균의 분포는 모집단의 원래 분포와 상관없이 정규분포에 가까워집니다.

적용 가능한 상황

- **이항 분포**: 제품의 합격/불합격, 클릭/비클릭, 구매/비구매 등 두 가지 결과만 갖는 사건의 발생 횟수를 모델링할 때.
- **포아송 분포**: 시간당 웹사이트 방문자 수, 하루 동안의 특정 교차로 사고 건수, 책 한 페이지의 오타 수 등 희귀 사건(rare event)의 발생 횟수를 모델링할 때.
- **정규 분포**: 사람들의 키나 몸무게, 시험 성적, 측정 오차 등 다양한 자연 및 사회 현상을 모델링하고, 통계적 가설 검정의 기반으로 사용될 때.

구현 방법

`scipy.stats` 모듈은 다양한 확률 분포에 대한 함수를 제공합니다. 주요 함수는 다음과 같습니다:

- `pmf` (Probability Mass Function): 이산 확률 변수의 특정 값에 대한 확률. (이항, 포아송)
- `pdf` (Probability Density Function): 연속 확률 변수의 특정 지점에서의 확률 밀도.
- `cdf` (Cumulative Distribution Function): 확률 변수가 특정 값보다 작거나 같을 확률.
- `rvs` (Random Variates): 해당 분포를 따르는 난수 생성.

이항 분포 (Binomial Distribution)

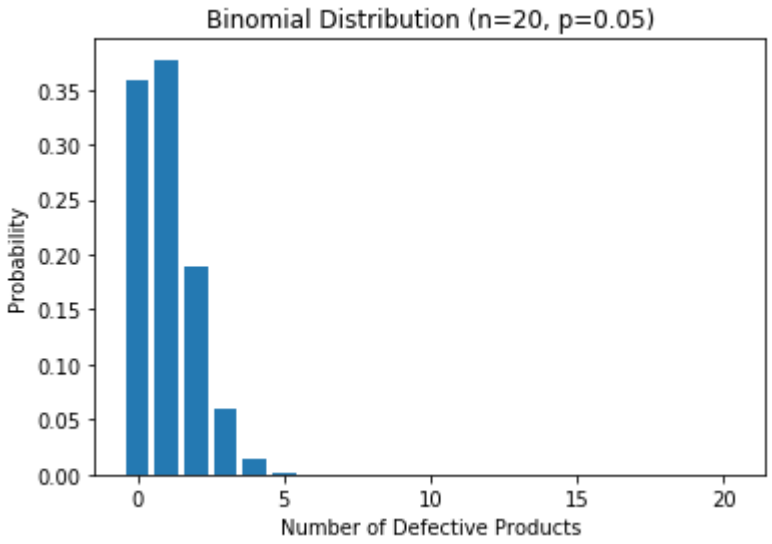
- **문제**: 불량률이 5%(p=0.05)인 제품 20개(n=20)를 검사할 때, 불량품이 정확히 2개(k=2) 나올 확률은?

```
from scipy.stats import binom
import matplotlib.pyplot as plt
import numpy as np

n, p = 20, 0.05
k = 2

# P(X=k) - pmf 사용
prob = binom.pmf(k, n, p)
print(f"불량품이 정확히 {k}개 나올 확률: {prob:.4f}") # 0.1887

# 시각화
x = np.arange(0, n+1)
plt.bar(x, binom.pmf(x, n, p))
plt.title(f"Binomial Distribution (n={n}, p={p})")
plt.xlabel("Number of Defective Products")
plt.ylabel("Probability")
plt.show()
```



포아송 분포 (Poisson Distribution)

- **문제**: 어떤 가게에 시간당 평균 10명(λ=10)의 손님이 온다. 앞으로 한 시간 동안 손님이 정확히 5명(k=5)올 확률은?

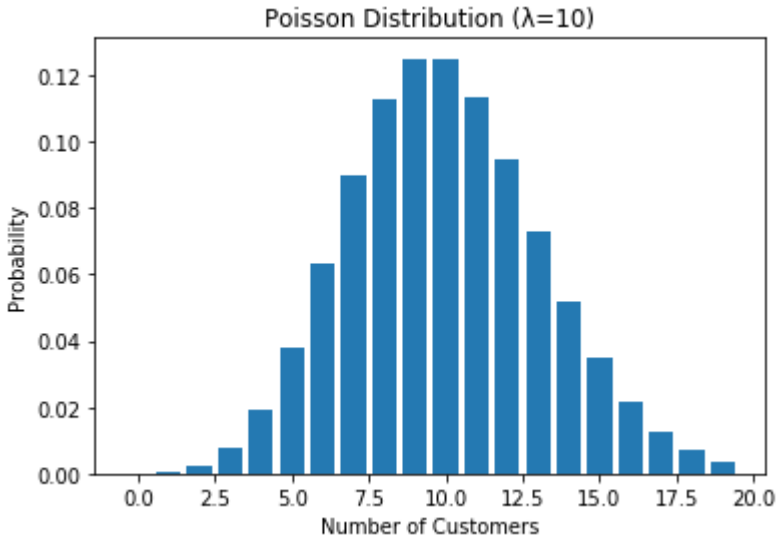
```
from scipy.stats import poisson

mu = 10 # lambda
k = 5

# P(X=k) - pmf 사용
```

```
prob = poisson.pmf(k, mu)
print(f"손님이 정확히 {k}명 올 확률: {prob:.4f}") # 0.0378

# 시각화
x = np.arange(0, 2 * mu)
plt.bar(x, poisson.pmf(x, mu))
plt.title(f"Poisson Distribution (λ={mu})")
plt.xlabel("Number of Customers")
plt.ylabel("Probability")
plt.show()
```



정규 분포 (Normal Distribution)

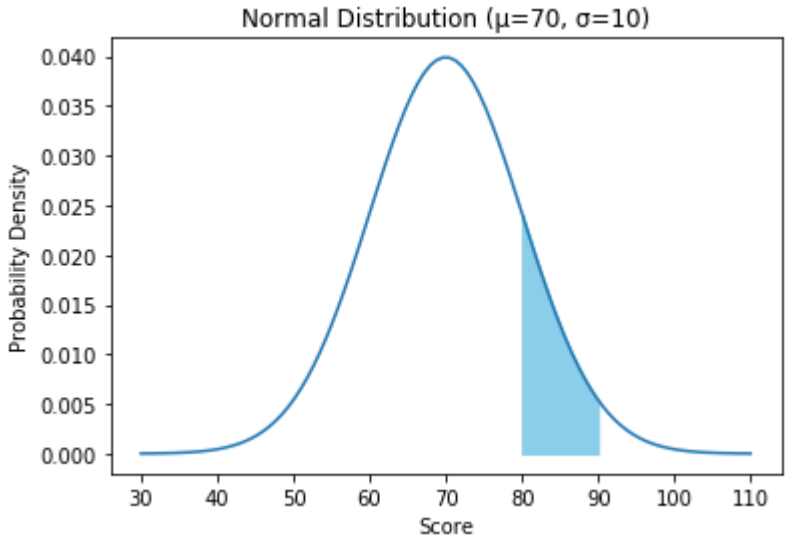
- **문제**: 어떤 시험의 성적은 평균이 70점(μ=70), 표준편차가 10점(σ=10)인 정규분포를 따른다. 점수가 80점 이상 90점 이하일 확률은?

```
from scipy.stats import norm

mu, sigma = 70, 10

# P(80 <= X <= 90) = P(X <= 90) - P(X < 80)
# cdf(x)는 P(X <= x)를 계산
prob = norm.cdf(90, loc=mu, scale=sigma) - norm.cdf(80, loc=mu, scale=sigma)
print(f"점수가 80점 이상 90점 이하일 확률: {prob:.4f}") # 0.1359

# 시각화
x = np.linspace(mu - 4*sigma, mu + 4*sigma, 100)
plt.plot(x, norm.pdf(x, loc=mu, scale=sigma))
plt.title(f"Normal Distribution (μ={mu}, σ={sigma})")
plt.xlabel("Score")
plt.ylabel("Probability Density")
# 확률 영역 채우기
x_fill = np.linspace(80, 90, 100)
plt.fill_between(x_fill, norm.pdf(x_fill, loc=mu, scale=sigma), color='skyblue')
plt.show()
```



분포 간의 관계

- **이항 분포와 정규 분포**: 이항 분포에서 시행 횟수 n이 충분히 크고, np와 $n(1-p)$가 모두 특정 값(보통 5 또는 10) 이상이면, 이항 분포는 평균 $μ=np$, 분산 $σ²=np(1-p)$인 정규분포에 근사합니다. 이는 계산이 복잡한 이항 분포를 정규분포로 근사하여 쉽게 계산할 수 있게 해줍니다.
- **포아송 분포와 이항 분포**: 이항 분포에서 n이 매우 크고 성공 확률 p가 매우 작을 때, 이항 분포는 $λ = np$인 포아송 분포에 근사합니다. 이는 포아송 분포가 '희귀 사건'의 분포로 사용되는 이유를 설명합니다.
- **포아송 분포와 정규 분포**: 포아송 분포에서 평균 발생 횟수 $λ$가 충분히 크면(보통 20 이상), 포아송 분포는 평균 $μ=λ$, 분산 $σ²=λ$인 정규분포에 근사합니다.