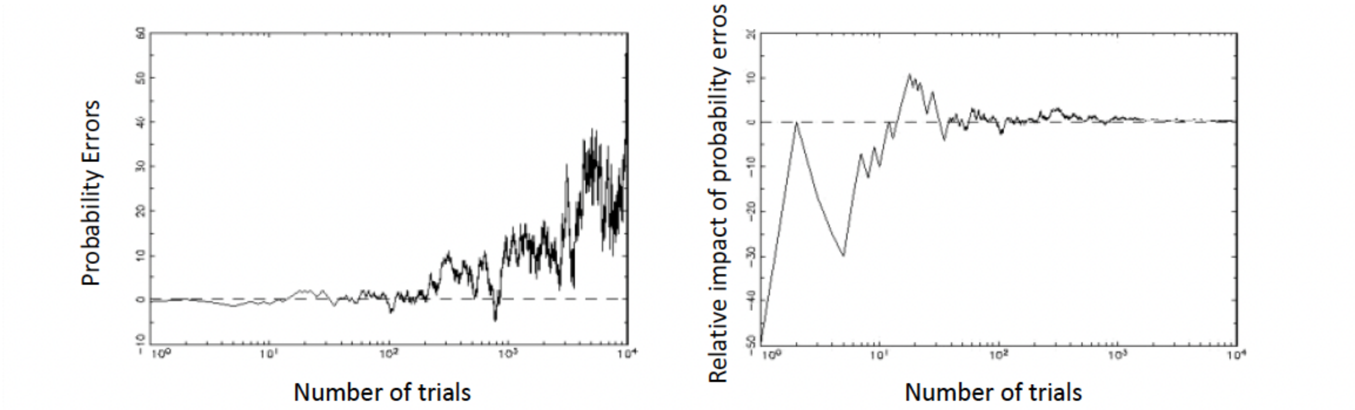
Probability Error, Law of Large Numbers, Expected Value, Standard Error, Binomial and Bernoulli Distributions

1. 확률 오류 (Probability Error)

- 확률 오류는 관찰된 결과와 예상된 확률 사이의 편차를 말한다. 예를 들어, 공정한 동전을 10번 던져서 나오는 결과(수학적 확률;예상은 5회 앞면, 5회 뒷면)와 실제 결과(실험적 확률;예를 들어 6회 앞면, 4회 뒷면) 사이의 차이를 확률 오류로 설명한다. 이러한 편차는 무작위 과정에서 자연스럽게 발생한다.

2. 큰 수의 법칙 (Law of Large Numbers)



- 큰 수의 법칙에 따르면, 시도 횟수를 늘릴수록 관찰된 결과의 상대 빈도(예: 동전 던지기에서 앞면이 나오는 비율)는 이론적 확률에 점점 가까워진다. 즉, 동전을 많이 던질수록 앞면과 뒷면이 나오는 비율은 각각 50%에 가까워지며, 확률 오류의 영향을 줄일 수 있다.

3. 기대값 (Expected Value)

- 기대값은 특정 확률 분포를 가진 임의의 변수에서 반복 실험을 무한히 많이 수행했을 때 나타날 것으로 예상되는 평균 결과이다. 이항 분포에서 기대값은 각 시행의 성공 확률을 모두 더한 값이며, 예를 들어, 6면의 확률이 공정한 주사위를 던질 때의 기대값은 3.5이다.

평균 vs. 기댓값

평균 (Mean)

- 통계적 맥락에서의 평균은 관찰된 데이터의 평균값을 나타낸다. 즉, 주어진 데이터 집합에서 모든 관측치를 더한 뒤, 관측치의 개수로 나눈 결과이다.
- 평균은 실제로 관측된 데이터를 바탕으로 계산되므로, 데이터가 주어져야만 그 값을 구할 수 있다.
- 예를 들어, 어떤 시험의 점수가 70, 80, 90점이 있다면, 이 점수들의 평균은 $(70 + 80 + 90) / 3 = 80$이다.

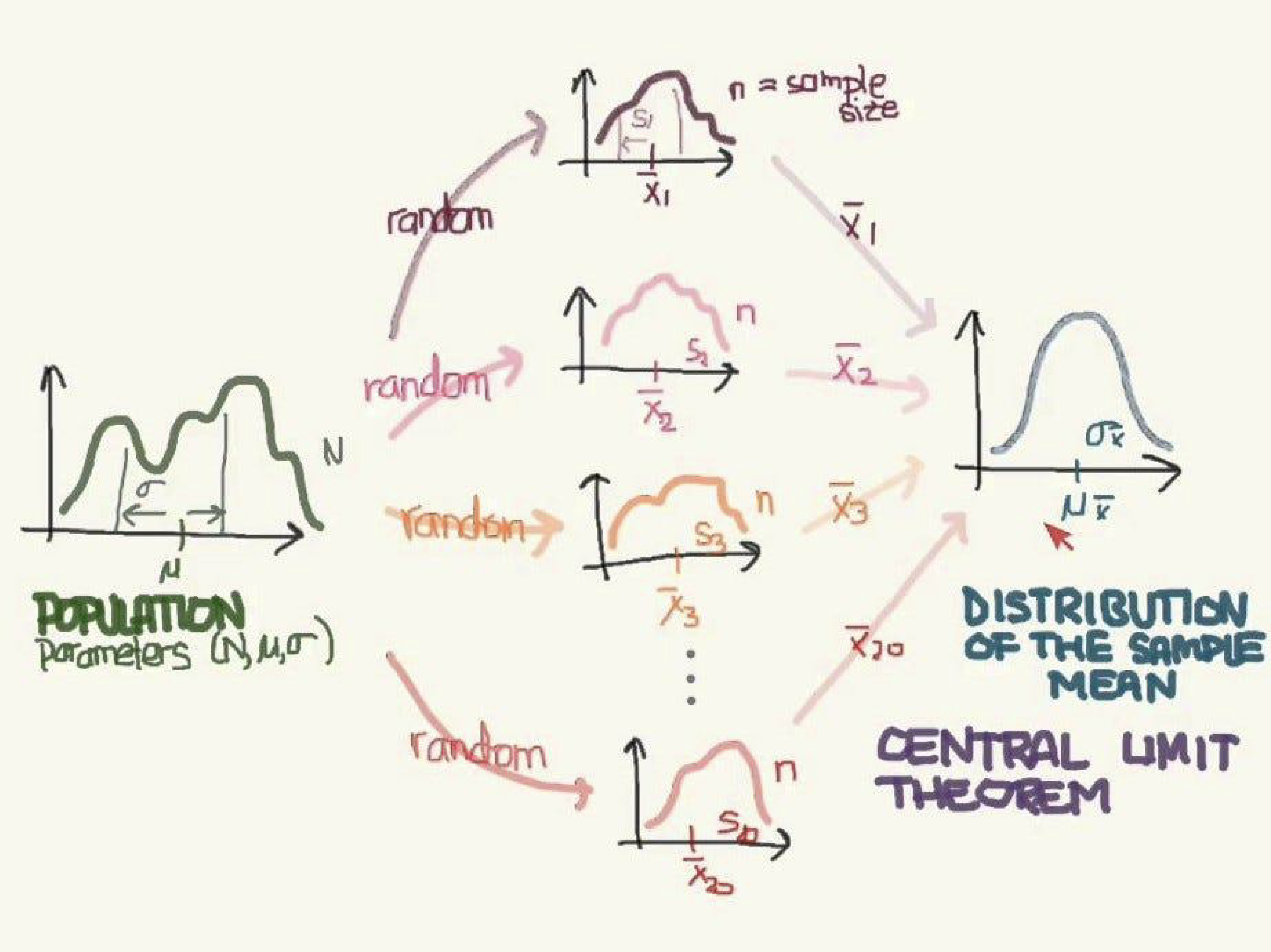
기대값 (Expected Value)

- 확률론적 맥락에서의 기대값은 임의의 확률 변수가 어떤 확률 분포를 따를 때, 무한히 많은 시행을 거쳐 나올 수 있는 평균적인 결과를 나타낸다.
- 기대값은 이론적인 개념으로, 확률 변수의 모든 가능한 값과 그 값이 발생할 확률을 고려하여 계산된다.
- 예를 들어, 공정한 주사위를 던졌을 때 나오는 수의 기대값은 $(1+2+3+4+5+6)/6 = 3.5$이다. 이 값은 실제로 주사위를 던져서 나올 수 있는 평균값을 이론적으로 나타낸다.

같아질때?

- 데이터가 특정 확률 분포를 따르고, 그 분포의 파라미터가 잘 정의되어 있을 때
- 데이터가 모집단 분포를 잘 대표하는 충분히 큰 샘플로부터 추출
- 편향이 없는 상황에서 무작위로 추출되었을 때

4. 표준 오차 (Standard Error)



- 표준 오차는 표본 통계량(예: 표본 평균)이 모집단 매개변수(예: 모집단 평균)에 대한 추정에서 나타날 수 있는 변동성 또는 불확실성을 측정한다. 표준 오차는 모집단의 표준편차를 표본 크기의 제곱근으로 나눈 값으로 계산된다. 이 값은 표본 평균이 모집단 평균 주변에서 얼마나 변동할지를 나타낸다.

- 표준편차(σ)

> $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

- x_i는 각 관측치의 값
- $\bar{x}$는 표본의 평균
- n은 표본의 크기
- σ는 표본의 표준편차

분모에 $n - 1$을 사용하는 이유는 표본 자유도를 고려하기 때문dl다. 표본 자유도를 사용하면 모집단 표준편차를 더 편향되지 않게 추정할 수 있다. 모집단 표준편차는 모든 모집단 구성원을 포함한 표준편차이며, 이는 일반

적으로 모집단의 실제 통계적 특성을 정확히 나타낸다. 반면, 여기서 언급한 표준편차는 특정 표본에서 계산된 값으로 모집단 표준편차의 추정치로 활용된다.

- 분산(σ^2)
- 표준오차($\sigma_{\bar{x}}$;SEM)

> $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

- $\bar{x}$: 표본 평균
- n: 표본 크기

SEM

모집단의 mean과 샘플링을 통해 얻은 어떤 표본집단의 mean간의 어떠한 차이를 보여준다.

eg.주사위 던지기

공정한 6면체 주사위를 던질 때, 각 면이 나올 확률은 $\frac{1}{6}$이고, 기대값(주사위의 평균값)은 $\frac{1+2+3+4+5+6}{6} = 3.5$이다. 각 면이 나오는 것은 독립적인 사건이고, 모든 면이 나올 확률이 같으므로, 분산(variance)은 각 면의 값에서 기대값을 뺀 다음 제곱하고, 이를 모두 더한 다음에 6으로 나눈 값이다.

$\text{Variance} = \frac{1}{6}\sum_{i=1}^{6} (i - 3.5)^2$

$\text{Standard Deviation} = \sqrt{\text{Variance}}$ 1.7078에 근접한 값을 얻을 수 있다. $\text{SEM} = \frac{\text{Standard Deviation}}{\sqrt{36}}$

실제 계산을 하면 주사위 던지기에서 나온 SEM 값은 표본 평균이 주사위 던지기의 진짜 평균인 3.5에서 기대되는 표준 편차를 나타낸다. 만약 36번의 주사위 던지기를 많이 반복한다면, 그 평균들은 약 0.2846의 표준 편차를 가진 분포를 형성할 것으로 기대할 수 있다. 이 값은 표본 평균의 분포가 얼마나 모집단의 진짜 평균 주위에 밀집해 있는지를 보여준다.

5. 베르누이 분포 (Bernoulli Distributions)

- 베르누이 분포는 단일 시도에서 성공(1) 또는 실패(0)의 이진 결과를 모델링하는 가장 간단한 분포이다. 특징은 파라미터로 p만을 갖고있다.(성공확률 p, 실패확률 1-p)

6. 이항 분포 (Binomial Distributions)

- 이항 분포는 베르누이 분포(Bernoulli Distribution)를 확장한 것으로 볼 수 있다. 베르누이 분포는 단일 시행에서의 두 가지 결과(성공 또는 실패)를 모델링하는 반면, 이항 분포는 여러 번의 독립적인 시행에서 나타나는 성공 횟수를 모델링한다. 즉, 이항 분포는 여러 베르누이 시행의 결과를 합산하여 하나의 분포로 표현한 것이다. 예를 들어, 동전을 10번 던져 앞면이 나오는 횟수를 모델링할 때 이항 분포를 사용한다. 두개의 파라미터를 사용한다.(성공확률 p, 시행횟수 n) eg. 동전던지기

$$k = 0 : \frac{10!}{0! \cdot 10!} \left(\frac{1}{2}\right)^0 \cdot \left(1 - \frac{1}{2}\right)^{10} = 0.0010$$

$$k = 1 : \frac{10!}{1! \cdot 9!} \left(\frac{1}{2}\right)^1 \cdot \left(1 - \frac{1}{2}\right)^{9} = 0.0098$$

$$k = 2 : \frac{10!}{2! \cdot 8!} \left(\frac{1}{2}\right)^2 \cdot \left(1 - \frac{1}{2}\right)^{8} = 0.0439$$

$$\vdots$$

$$k = 10 : \frac{10!}{10! \cdot 0!} \left(\frac{1}{2}\right)^{10} \cdot \left(1 - \frac{1}{2}\right)^{0} = 0.0010$$

계산

한번 시행에 대한 확률 구하기(베르누이)

$P(X_1 = x) = p^x (1 - p)^{1-x}$ $E(X_1) = \sum_{x} xP(X_1 = x) = 0 \times (1-p) + 1 \times p = p$ $\text{Var}(X_1) = E((X_1 - E(X_1))^2) = \sum_{x} (x - p)^2 P(X_1 = x) = (0 - p)^2 (1-p) + (1 - p)^2 p = p(1 - p)$

각각의 시행이 동일한 확률을 가지고 있다는 가정을 하고 진행한다(당연한 이야기)

- 기대값 (Expected Value) X의 기대값 $E(X)$는 n개의 독립적인 베르누이 시행의 성공 횟수의 합이다. 각 베르누이 시행 $X_1, X_2, \ldots, X_n$의 기대값은 p이므로, 전체 기대값은: $E(X) = E(X_1 + X_2 + \ldots + X_n) = nE(X_1) = np$
- 분산 X의 분산 $\text{Var}(X)$는 n개의 독립적인 베르누이 시행의 성공 횟수의 분산의 합입니다. 각 베르누이 시행의 분산은 $p(1-p)$이므로, 전체 분산은: $\text{Var}(X) = \text{Var}(X_1 + X_2 + \ldots + X_n) = n\text{Var}(X_1) = np(1-p)$

예시

동전을 한 번 던질 때:

1. **확률(p)**: 앞면이 나올 확률은 0.5 입니다.
2. **기대값(E[X1])**: $p = 0.5$
3. **분산(Var[X1])**: $p(1-p) = 0.5 \times (1-0.5) = 0.25$

동전을 100번 던질 때:

1. **기대값(E[X])**: $n \times p = 100 \times 0.5 = 50$
2. **분산(Var[X])**: $n \times p \times (1-p) = 100 \times 0.5 \times 0.5 = 25$