# 선수 지식 - 통계
## 평균과 기댓값

평균과 기댓값 | 딥러닝의 기초가 되는 확률 개념 알아보기

강사 나동빈

# 선수 지식 - 통계

## 평균과 기댓값

# 평균(Mean)



- 평균에는 다양한 종류가 있으며, 가장 일반적인 평균은 산술 평균이다.
- **산술 평균**(arithmetic mean): 모든 관측 값을 더해 관측 값의 개수로 나눈 것이다.

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots a_n}{n}$$

# 평균(Mean) 계산 예시



- 5명의 학생의 딥러닝 강의 중간고사 점수가 각각 56, 93, 88, 72, 65점이다.
- **점수의 평균**: (56 + 93 + 88 + 72 + 65) / 5 = 74.80

| 학생 번호 | 1번 | 2번 | 3번 | 4번 | 5번 | 평균 |
|---|---|---|---|---|---|---|
| 성적 | 56 | 93 | 88 | 72 | 65 | 74.80 |

# 특정한 집단을 대표하는 값



- **평균**(mean)은 특정한 데이터 집단을 대표하기에 적절한가?

→ 다음의 사례를 확인해 보자.

**[예시]** 미국의 노스캐롤라니아 대학의 졸업생 평균 연봉이 가장 높은 학과는?

→ 전문직 종사자가 많은 학과가 아닌, 지리학과가 평균 연봉 1억 이상으로 1등을 한 적이 있다.

- 당시 마이클 조던이 해당 학교 지리학과 졸업생 중 하나였기 때문이다.

# 중앙값(Median)

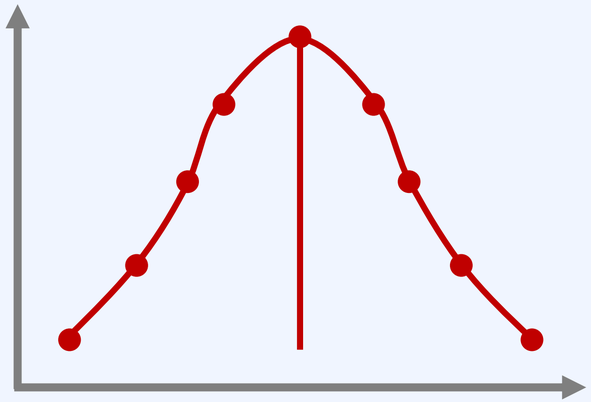


- **중앙값(median)**: 주어진 값들을 순서대로 정렬했을 때, 가장 중앙에 위치하는 값이다.
- 예를 들어 3, 5, 100이 있을 때 평균은 36이지만, 중앙값은 5이다.

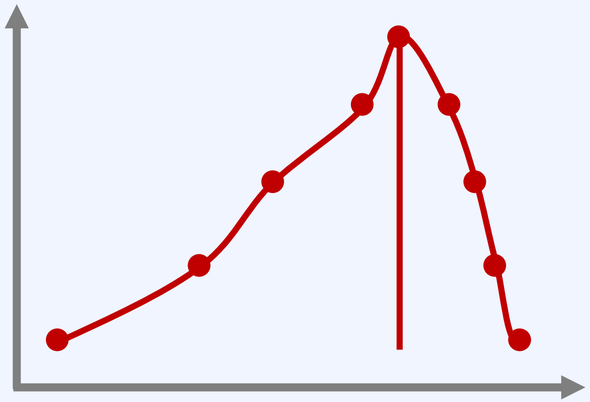
# 평균(Mean) vs. 중앙값(Median)



- **평균**(mean)과 **중앙값**(median)은 각각 어떤 상황에서 효과적일까?
- **평균**: 데이터의 분포가 정규분포처럼 대칭적인 경우
- **중앙값**: 데이터의 분포가 한쪽으로 치우쳐졌거나(skewed), 이상치(outlier)가 존재하는 경우



대칭 분포



비대칭 분포

# 평균(Mean) vs. 중앙값(Median)



- 예를 들어 이상치(outlier)가 존재하는 상황을 생각해 보자.
- 6명의 학생의 딥러닝 강의 중간고사 점수가 각각 56, 93, 88, 72, 65, 2점이다.
- **점수의 평균**: (56 + 93 + 88 + 72 + 65 + 2) / 5 = 62.67

→ 하나의 값에 의하여 평균 값이 급격히 떨어진다.

| 번호 | 1번 | 2번 | 3번 | 4번 | 5번 | 6번 | 평균 |
|---|---|---|---|---|---|---|---|
| 성적 | 56 | 93 | 88 | 72 | 65 | 2 | 62.67 |

# 평균(Mean) vs. 중앙값(Median)



- 평균(mean)과 중앙값(median)에 대한 설명은 다음의 표에서 확인할 수 있다.

| | 평균(mean) | 중앙값(median) |
|---|---|---|
| 사용 사례 | 일반적으로 정규 분포를 따르는 데이터에 사용된다. | 일반적으로 비대칭 분포(skewed distribution)에서 사용된다. |
| 특징 | 이상치(outlier)에 영향을 많이 받아 값이 크게 변할 수 있다. | 이상치(outlier)에 대하여 강건한(robust) 값의 형태를 띤다. |
| 계산 방법 | 모든 값을 더한 뒤에 값의 개수로 나누기 | 정렬 이후에 중간에 위치한 값 채택 |

# 기댓값(Expectation)



- 각 사건에 대해 **확률 변수와 확률 값을 곱하여**, 전체 사건에 대하여 모두 더한 값이다.
- $E[X] = \sum_i \{(i\text{번째 사건이 발생할 확률}) \times (i\text{번째 사건에 대한 확률 변수})\}$
- 사실상 기댓값은 산술 평균과 유사하며, 실제로 두 용어를 섞어서 사용하곤 한다.

# 기댓값(Expectation)



- 기댓값은 모든 사건에 대해 확률을 곱하면서 더하여 계산할 수 있다.
- 이산확률변수에 대한 기댓값은 다음의 공식을 통해 계산할 수 있다.

$$E[X] = \sum_i x_i \cdot f(x_i)$$

- 연속확률변수에 대한 기댓값은 다음의 공식을 통해 계산할 수 있다.

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

$\boldsymbol{x}$: 사건

$\boldsymbol{f(x)}$: 확률 분포 함수

# 평균(Mean) vs. 기댓값(Expectation)



- 평균과 기댓값은 그 의미가 유사하지만, 일반적으로 사용되는 문맥이 다르다.
- **기댓값**: 새로운 데이터가 관측되었을 때, 그 데이터가 확률적으로 어떤 값을 가질지를 예측할 때
- **평균**: 이미 구해진 값에 대하여 통계적인 특성을 분석할 때

# 기댓값(Expectation) – 복권 기대 이익 문제



- 1,000원짜리 복권이 있다.

| 등급 | 당첨 확률 | 당첨금 |
|---|---|---|
| 1등 | 1 / 10 | 5,000원 |
| 2등 | 3 / 10 | 1,000원 |
| 꽝 | 6 / 10 | 0원 |

- 이때 복권을 사서 얻을 수 있는 **기대 이익(수익 – 비용)**은?

# 기댓값(Expectation) – 복권 기대 이익 문제



- **당첨 금액**을 확률 변수 $X$라고 해보자.
- 각 당첨 금액은 이산적인 형태를 가지기 때문에, 이산확률변수로 볼 수 있다.
- 확률 값은 다음과 같이 표현 가능하다.
- $P[X = 5000] = 1/10$
- $P[X = 1000] = 3/10$
- $P[X = 0] = 6/10$

# 기댓값(Expectation) – 복권 기대 이익 문제



- 따라서 당첨금에 대한 기댓값은 다음과 같다.

→ $(5{,}000 \times 1/10) + (1{,}000 \times 3/10) + (0 \times 6/10) = 800$원

- 이때 복권의 가격이 1,000원이므로, 기대 이익은 -200원이다.
- 따라서 복권을 사지 않는 것이 이득이다.

# 기댓값(Expectation) – 나이 예측 모델



- 특정한 사람 얼굴 이미지 $x$가 주어진 상황을 생각해 보자.
- 얼굴의 나이를 예측하는 모델 $f(x)$는 10세부터 13세 범위의 나이에 대해 확률을 예측한다.
- 해당 모델이 예측한 결과 $P(y|x)$는 다음과 같다.

| $y$ | 10 | 11 | 12 | 13 |
|---|---|---|---|---|
| $P(y\|x)$ | 0.2 | 0.4 | 0.3 | 0.1 |

- 이때 **기댓값**을 계산한다면?

→ $(10 \times 0.2) + (11 \times 0.4) + (12 \times 0.3) + (13 \times 0.1) = 11.3$