# Part 3,4 - 요약 및 문제

## Part 3. Working with Data

### 기술통계 분석 지표

기술통계 분석은 데이터의 기본적인 특성을 파악하는 데 사용되는 다양한 지표들로 구성됩니다. 이러한 지표들은 데이터의 중심 경향, 분산, 비대칭성 등을 측정하여 데이터의 특성을 요약하고 이해하는 데 도움을 줍니다.

### 1. 중심 경향 지표

- 평균(Mean): 모든 데이터 값의 합을 데이터 개수로 나눈 값
- 중앙값(Median): 데이터를 크기 순으로 정렬했을 때 가운데 값
- 최빈값(Mode): 가장 자주 나타나는 데이터 값

### 2. 분산 지표

- 범위(Range): 최댓값 - 최솟값
- 사분위간 범위(IQR): 제3사분위수 - 제1사분위수
- 평균 절대 편차(MAD): Σ|x - mean(x)| / n
- 분산(Variance): Σ(x - mean(x))^2 / n
- 표준편차(Standard Deviation): sqrt(Variance)
- 중앙 절대 편차(MAD): median(|x - median(x)|)

### 3. 비대칭성 지표

- 왜도(Skewness): Σ(x - mean(x))^3 / (n * std(x)^3)
- 첨도(Kurtosis): Σ(x - mean(x))^4 / (n * std(x)^4) - 3

### 각각의 공식과 간단예시

- **평균(Mean)**

+ 공식: Σx / n, 여기서 Σx는 모든 데이터 값의 합이고 n은 데이터 개수

+ 예시) 데이터 [2, 4, 6, 8, 10]의 평균 = (2 + 4 + 6 + 8 + 10) / 5 = 6

- **중앙값(Median)**

+ 공식: 데이터를 크기 순으로 정렬했을 때 가운데 값

+ 예시) 데이터 [2, 4, 6, 8, 10]의 중앙값 = 6

- **최빈값(Mode)**

+ 공식: 가장 자주 나타나는 데이터 값

+ 예시) 데이터 [2, 4, 4, 6, 8, 10]의 최빈값 = 4

- **범위(Range)**

+ 공식: 최댓값 - 최솟값

+ 예시) 데이터 [2, 4, 6, 8, 10]의 범위 = 10 - 2 = 8

- **사분위간 범위(IQR)**

+ 공식: 제3사분위수 - 제1사분위수

+ 예시) 데이터 [2, 4, 6, 8, 10]의 IQR = 8 - 4 = 4

- **평균 절대 편차(MAD)**

+ 공식: Σ|x - mean(x)| / n

+ 예시) 데이터 [2, 4, 6, 8, 10]의 MAD = (|2-6| + |4-6| + |6-6| + |8-6| + |10-6|) / 5 = 2.4

- **분산(Variance)**

+ 공식: Σ(x - mean(x))^2 / n

+ 예시) 데이터 [2, 4, 6, 8, 10]의 분산 = ((2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2) / 5 = 8

- **표준편차(Standard Deviation)**

+ 공식: sqrt(Variance)

+ 예시) 데이터 [2, 4, 6, 8, 10]의 표준편차 = sqrt(8) = 2.83

- **중앙 절대 편차(MAD)**

+ 공식: median(|x - median(x)|)

+ 예시) 데이터 [2, 4, 6, 8, 10]의 MAD = median(|2-6|, |4-6|, |6-6|, |8-6|, |10-6|) = 2

- **왜도(Skewness)**

+ 공식: Σ(x - mean(x))^3 / (n * std(x)^3)

+ 예시) 데이터 [2, 4, 6, 8, 10]의 왜도 = 0 (완전 대칭)

- **첨도(Kurtosis)**

+ 공식: Σ(x - mean(x))^4 / (n * std(x)^4) - 3

+ 예시) 데이터 [2, 4, 6, 8, 10]의 첨도 = -1.2 (분포가 정규분포보다 가벼운 편)

### 문제) 어떤 회사의 직원들의 연봉 데이터는 다음과 같습니다.

이 데이터의 평균과 표준편차를 계산하고 히스토그램으로 시각화 하시오

연봉 데이터: 3000, 3500, 4000, 4500, 5000, 5500, 6000

```
import numpy as np
import matplotlib.pyplot as plt

salaries = [3000, 3500, 4000, 4500, 5000, 5500, 6000]

# 평균, 표준편차 구하는 코드를 적으시오


# 히스토그램으로 데이터를 시각화 하는 코드를 적으시오
```

# Part 4. Statistical Theory

## T-Test / ANOVA 가설검증 예시 시나리오 및 방법

## 용어 설명

- **t-test 가설검증**

t-test는 두 집단의 평균 차이가 통계적으로 유의한지 검정하는 방법입니다. 주로 다음과 같은 상황에서 사용됩니다.

두 집단의 평균 차이 검정특정 집단의 평균이 특정 값과 다른지 검정대응 표본(paired samples) 간 평균 차이 검정

t-test 결과 해석 시, p-value가 유의수준(일반적으로 0.05) 보다 작으면 귀무가설을 기각하고 대립가설을 채택합니다.

이는 두 집단의 평균이 통계적으로 유의한 차이가 있다는 것을 의미합니다.

- **ANOVA 가설 검증**

ANOVA(Analysis of Variance)는 세 개 이상의 집단 평균 차이를 검정하는 방법입니다.

주로 다음과 같은 상황에서 사용됩니다.

세 개 이상의 집단 평균 차이 검정독립변수가 2개 이상인 경우의 집단 간 평균 차이 검정

ANOVA 결과 해석 시, p-value가 유의수준(일반적으로 0.05) 보다 작으면 귀무가설을 기각하고 대립가설을 채택합니다.

이는 집단 간 평균의 차이가 통계적으로 유의하다는 것을 의미합니다.

ANOVA에서 귀무가설이 기각되면, 어떤 집단 간 차이가 있는지 확인하기 위해 사후 검정(post-hoc test)을 수행합니다. 대표적인 사후 검정 방법으로는 Tukey's HSD, Bonferroni 등이 있습니다.

이와 같이 t-test와 ANOVA는 서로 다른 상황에 적용되는 가설 검정 방법이지만, 공통적으로 귀무가설과 대립가설을 정의하고 통계적 유의성을 판단하는 원리를 따릅니다.

- **귀무가설 (H0):**

실험 결과에 차이가 없다는 가설. 즉, 두 집단 간 평균의 차이가 없거나, 세 집단 이상의 평균이 모두 같다는 가설.

- **대립가설 (H1):**

실험 결과에 차이가 있다는 가설. 즉, 두 집단 간 평균의 차이가 있거나, 세 집단 이상의 평균이 모두 같지 않다는 가설.

⁂ 가설 설정 시 실험의 목적과 연구 문제에 맞게 귀무가설과 대립가설을 정의합니다. 일반적으로 귀무가설은 "차이가 없다"는 보수적인 가설이며, 대립가설은 "차이가 있다"는 가설입니다.

## T-Test & ANOVA 예제 시나리오 설정

- **t-test 가설 검증**

```
import numpy as np
from scipy.stats import ttest_ind

# 데이터 셋
existing_product_scores = [4, 5, 3, 4, 5, 3, 4, 5, 4, 5]
new_product_scores = [5, 4, 4, 5, 3, 4, 5, 4, 5, 4]

# 가설 설정
# 귀무가설 (H0): 기존 제품과 신제품의 만족도 평균 차이가 없다.
# 대립가설 (H1): 기존 제품과 신제품의 만족도 평균 차이가 존재한다.

# 유의수준 설정
alpha = 0.05

# 검정통계량 계산
existing_mean = np.mean(existing_product_scores)
new_mean = np.mean(new_product_scores)
existing_std = np.std(existing_product_scores, ddof=1)
new_std = np.std(new_product_scores, ddof=1)
t_stat, p_value = ttest_ind(existing_product_scores, new_product

# 결과 해석
print(f"기존 제품 평균 점수: {existing_mean:.1f}")
print(f"신제품 평균 점수: {new_mean:.1f}")
print(f"t-값: {t_stat:.2f}")
print(f"p-값: {p_value:.4f}")

if p_value > alpha:
    print("귀무가설 H0을 기각할 수 없습니다.")
    print("기존 제품과 신제품의 만족도 평균 차이가 통계적으로 유의미하지 않
```

```
else:
    print("귀무가설 H0을 기각합니다.")
    print("기존 제품과 신제품의 만족도 평균 차이가 통계적으로 유의미합니다.'
```

```
기존 제품 평균 점수: 4.2
신제품 평균 점수: 4.4
t-값: 0.63
p-값: 0.5348
귀무가설 H0을 기각할 수 없습니다.
기존 제품과 신제품의 만족도 평균 차이가 통계적으로 유의미하지 않습니다.
```

- **ANOVA 가설 검증**

```
import numpy as np
from scipy.stats import f_oneway

# 데이터 생성
product_A = [85, 82, 88, 90, 87]
product_B = [78, 81, 75, 79, 82]
product_C = [92, 88, 90, 85, 89]

# 귀무가설과 대립가설 설정
# 귀무가설 (H0): 세 제품의 고객 만족도 평균은 모두 같다.
# 대립가설 (H1): 적어도 한 제품의 고객 만족도 평균이 다르다.
# ANOVA 실행
f_stat, p_value = f_oneway(product_A, product_B, product_C)

# 결과 출력
print(f"F-statistic: {f_stat:.2f}")
print(f"p-value: {p_value:.4f}")

# 가설 검정
alpha = 0.05
if p_value < alpha:
    print("귀무가설 기각, 적어도 한 제품의 고객 만족도 평균이 다르다.")
```

```
else:
    print("귀무가설 채택, 세 제품의 고객 만족도 평균은 모두 같다.")
```

```
F-statistic: 16.66
p-value: 0.0003
귀무가설 기각, 적어도 한 제품의 고객 만족도 평균이 다르다.
```