1. **균등 분포 데이터 (Uniform Distribution)**: 모든 값이 거의 동일한 확률로 나타납니다.
2. **정규 분포 데이터 (Normal Distribution)**: 평균 주변에 값들이 모여 있으며, 양쪽 꼬리로 갈수록 값이 줄어듭니다.
3. **이상치를 포함한 데이터 (Data with Outliers)**: 대부분의 값이 특정 범위 내에 있지만, 일부 이상치가 존재합니다.

각 데이터셋에 대해 정규화와 표준화를 적용하고, 결과를 도표로 나타내어 비교해 보겠습니다.

## 1. 정규화 (Normalization)

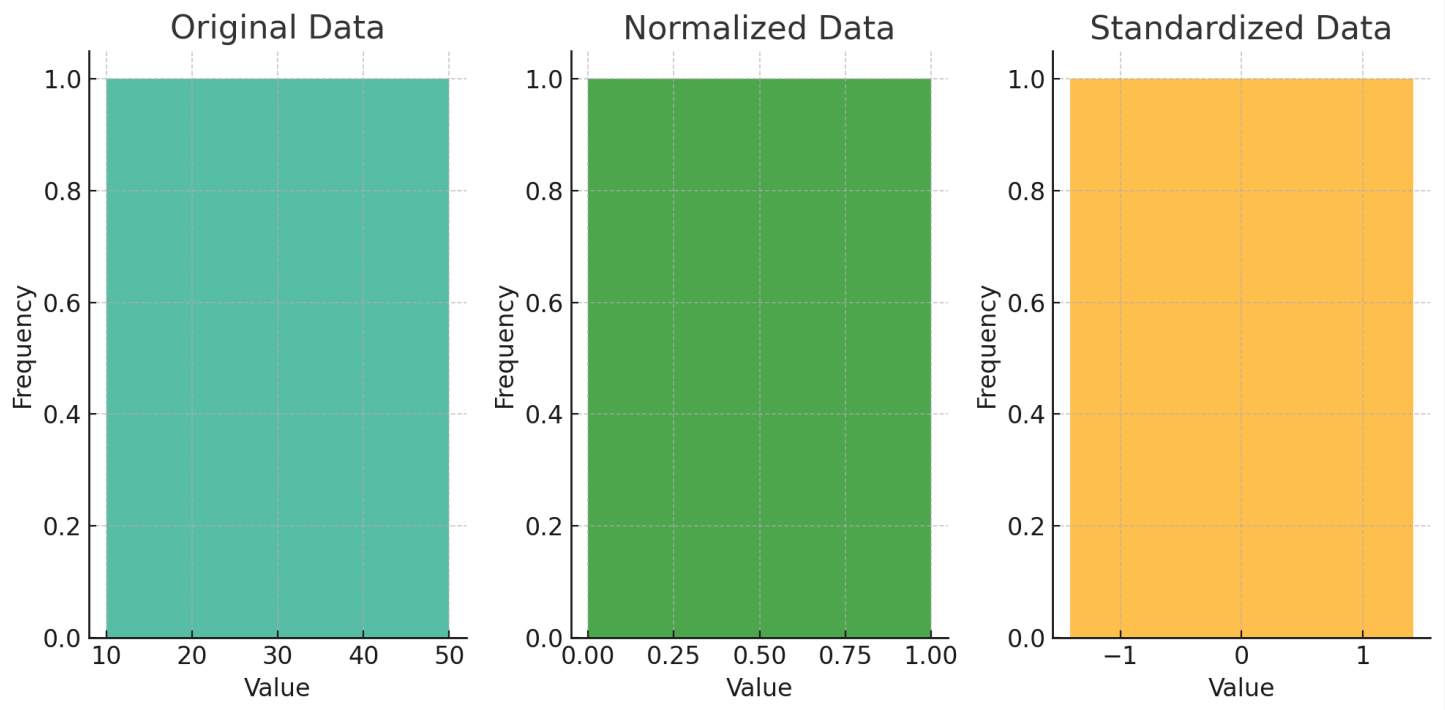
정규화는 데이터의 범위를 [0, 1]로 조정하는 과정입니다. 다음 공식을 사용합니다:

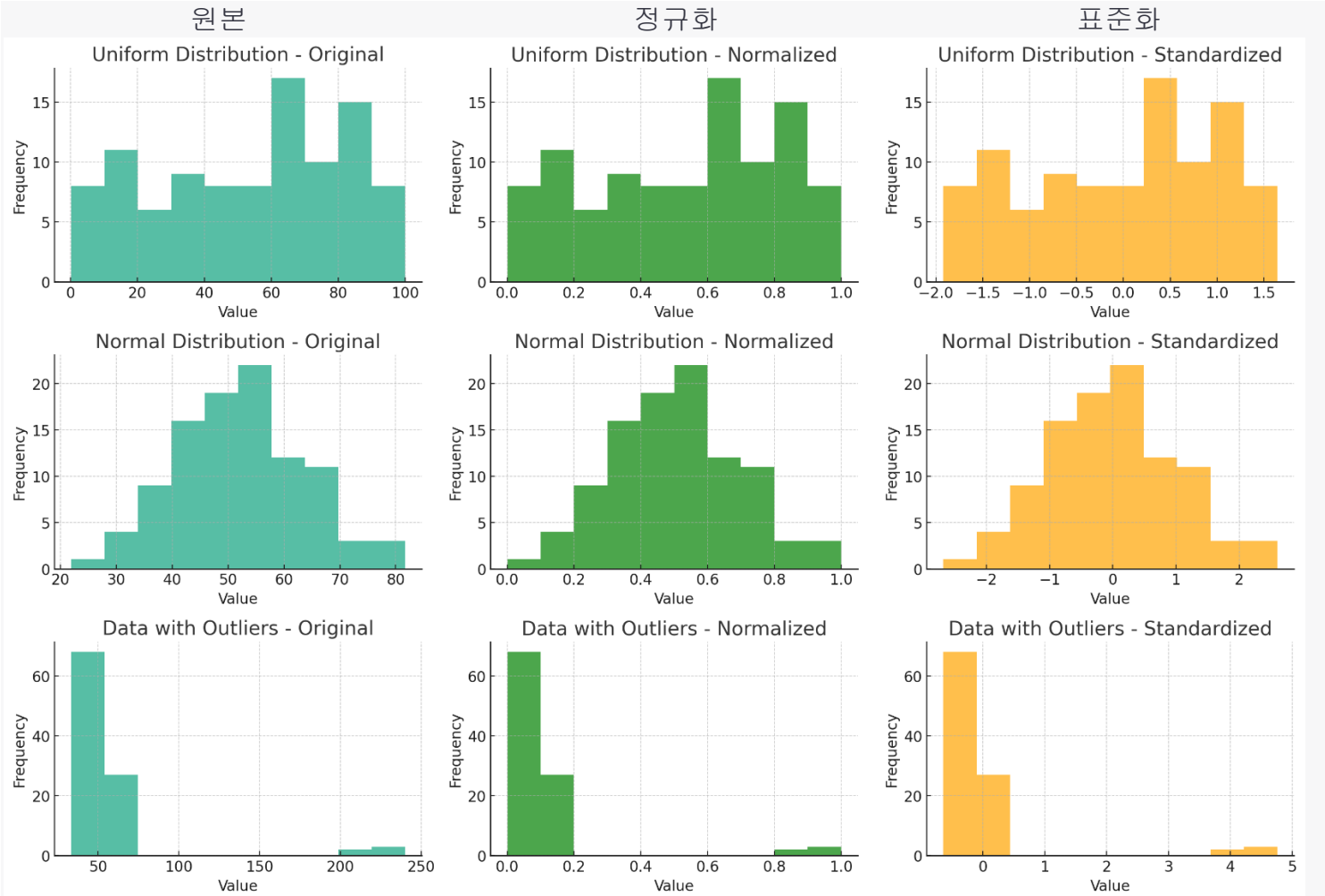
$$\text{Normalized} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

## 2. 표준화 (Standardization)

표준화는 데이터의 평균을 0, 표준편차를 1 로 조정하는 과정입니다. 다음 공식을 사용합니다:

$$\text{Standardized} = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$






도표를 통해 균등 분포 데이터, 정규 분포 데이터, 그리고 이상치를 포함한 데이터에 대해 정규화와 표준화가 어떻게 적용되는지 확인할 수 있습니다.

1. **균등 분포 데이터 (Uniform Distribution)**
   - 정규화: 값의 범위가 [0, 1]로 조정됩니다. 데이터의 상대적인 위치는 유지됩니다.
   - 표준화: 데이터의 평균이 0 주변으로 이동하고, 표준편차에 따라 값들이 분포합니다.
2. **정규 분포 데이터 (Normal Distribution)**
   - 정규화: 값의 범위가 [0, 1]로 조정됩니다. 데이터의 모양과 상대적인 위치는 유지됩니다.
   - 표준화: 데이터의 평균이 0 이 되고, 표준편차에 따라 값들이 분포합니다. 정규 분포의 모양은 유지됩니다.
3. **이상치를 포함한 데이터 (Data with Outliers)**
   - 정규화: 이상치로 인해 대부분의 데이터가 매우 좁은 범위 내에 압축됩니다. 이상치가 데이터의 스케일을 지배합니다.
   - 표준화: 이상치의 영향을 상대적으로 덜 받으며, 데이터의 분포가 더 잘 보존됩니다.

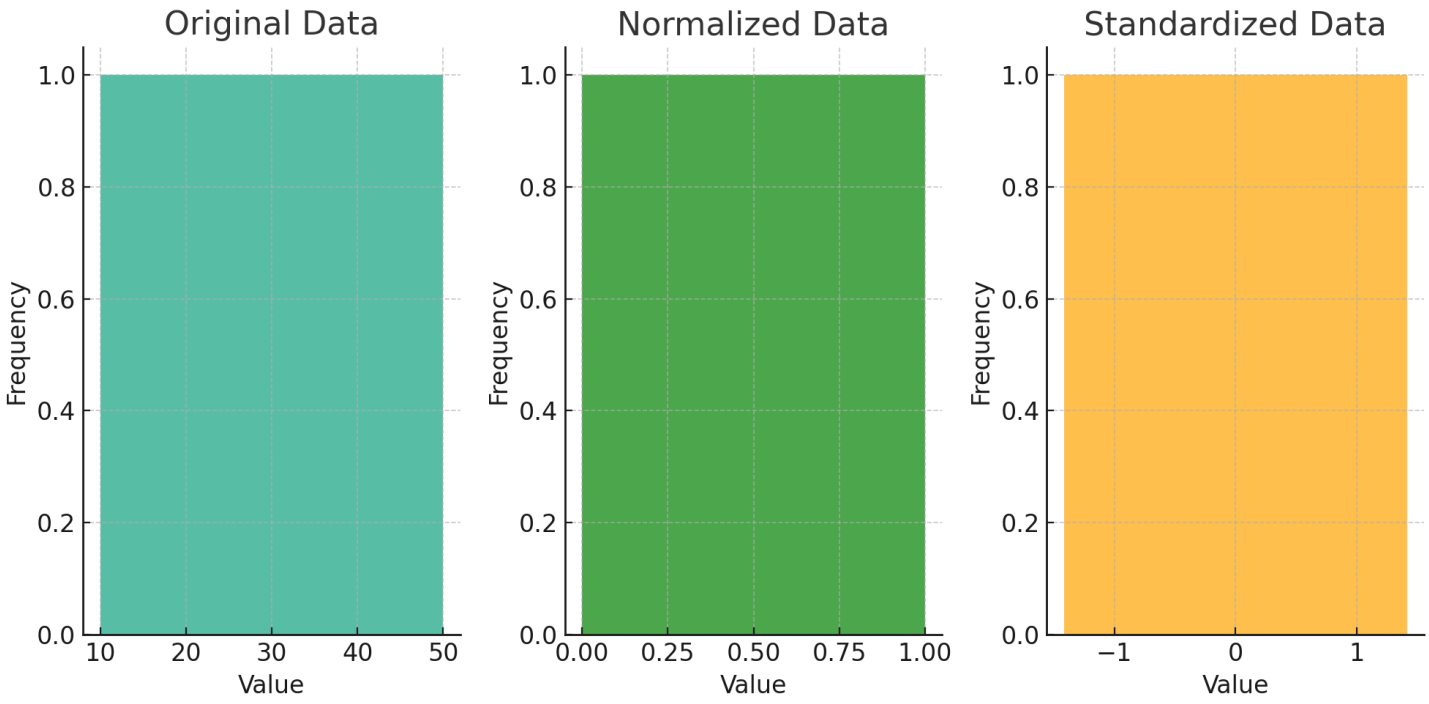
이러한 비교를 통해, 데이터의 종류와 목적에 따라 정규화와 표준화 중 어떤 방법을 선택할지 결정할 수 있습니다. 예를 들어, 이상치가 중요한 역할을 하는 경우에는 정규화보다는 표준화가 더 적합할 수 있습니다.

# 1. R 코드

```
# 데이터셋 생성
data <- c(10, 20, 30, 40, 50)

# 정규화
normalized_data <- (data - min(data)) / (max(data) - min(data))

# 표준화
standardized_data <- (data - mean(data)) / sd(data)

# 결과 출력
print("정규화된 데이터:")
print(normalized_data)

print("표준화된 데이터:")
print(standardized_data)
```

```
# 필요한 라이브러리 로드
library(ggplot2)

# 데이터셋 생성
set.seed(123)
data <- data.frame(
  Original = rnorm(100, mean = 50, sd = 10)
)

# 정규화
data$Normalized <- (data$Original - min(data$Original)) / (max(data$Original) - min(data$Original))

# 표준화
data$Standardized <- (data$Original - mean(data$Original)) / sd(data$Original)

# 도표 생성
ggplot(data) +
  geom_histogram(aes(x = Original, y = ..density..), binwidth = 1, fill = "blue", alpha = 0.5) +
  geom_histogram(aes(x = Normalized, y = ..density..), binwidth = 0.05, fill = "green", alpha = 0.5) +
  geom_histogram(aes(x = Standardized, y = ..density..), binwidth = 0.5, fill = "orange", alpha = 0.5) +
  labs(title = "Original, Normalized, and Standardized Data Distributions", x = "Value", y = "Density") +
  theme_minimal()
```

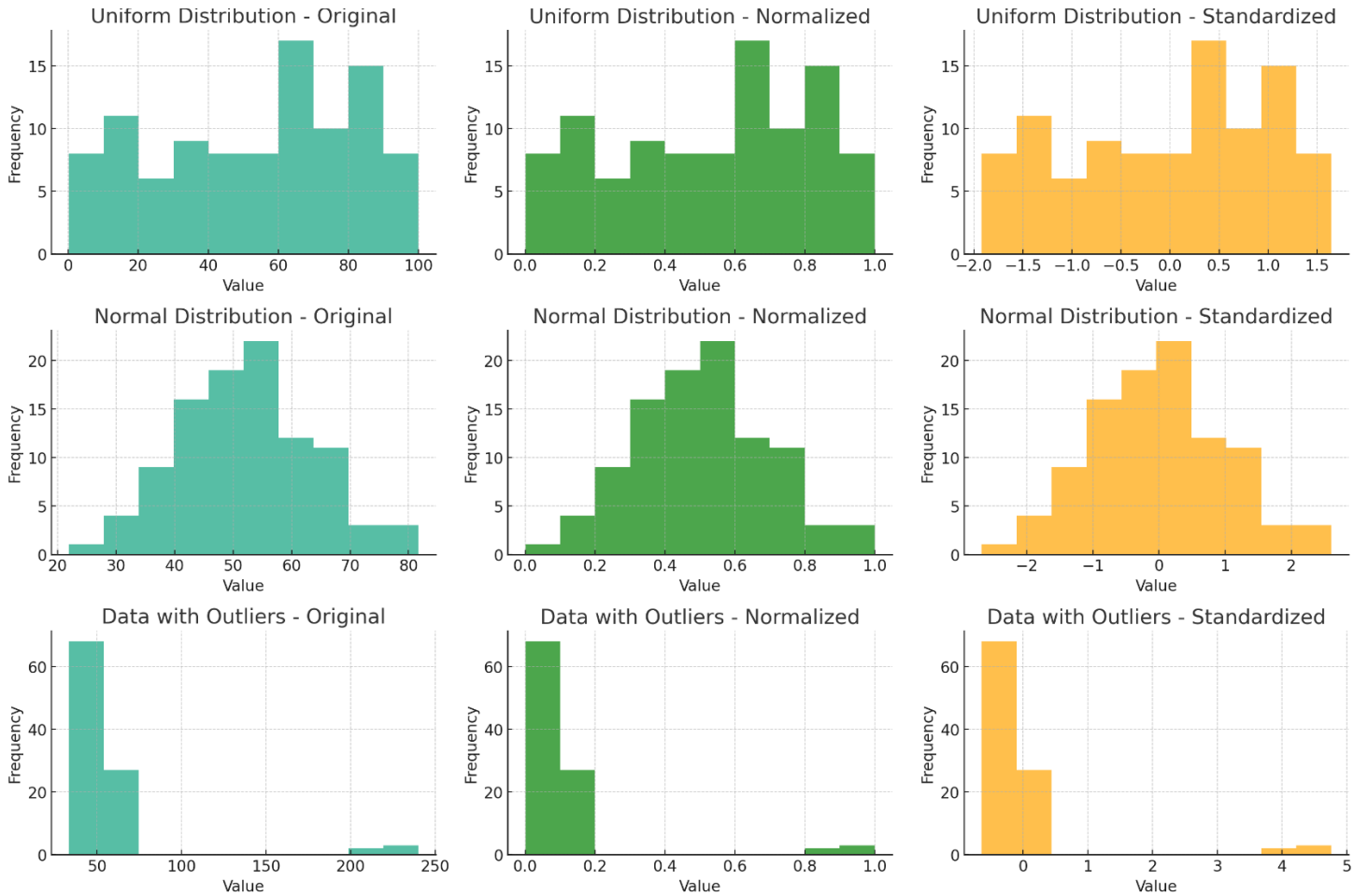
## 2. 파이썬 코드




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

# 데이터셋 생성
data = np.array([10, 20, 30, 40, 50])

# 정규화
normalized_data = (data - np.min(data)) / (np.max(data) - np.min(data))

# 표준화
standardized_data = (data - np.mean(data)) / np.std(data)

# 도표 생성
plt.figure(figsize=(10, 5))

# 원본 데이터 도표
plt.subplot(1, 3, 1)
plt.hist(data, bins=5, alpha=0.7, label='Original')
plt.title("Original Data")
plt.xlabel("Value")
plt.ylabel("Frequency")

# 정규화된 데이터 도표
plt.subplot(1, 3, 2)
```

```
plt.hist(normalized_data, bins=5, alpha=0.7, label='Normalized', color='green')
plt.title("Normalized Data")
plt.xlabel("Value")
plt.ylabel("Frequency")

# 표준화된 데이터 도표
plt.subplot(1, 3, 3)
plt.hist(standardized_data, bins=5, alpha=0.7, label='Standardized', color='orange')
plt.title("Standardized Data")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.tight_layout()
plt.show()
```

---




```
# 데이터셋 생성
uniform_data = np.random.uniform(low=0, high=100, size=100)
normal_data = np.random.normal(loc=50, scale=10, size=100)
outlier_data = np.random.normal(loc=50, scale=10, size=95)
outlier_data = np.append(outlier_data, [200, 210, 220, 230, 240])  # 이상치 추가

# 정규화 함수
```

```
def normalize(data):
    return (data - np.min(data)) / (np.max(data) - np.min(data))

# 표준화 함수
def standardize(data):
    return (data - np.mean(data)) / np.std(data)

# 데이터 정규화 및 표준화
uniform_normalized = normalize(uniform_data)
uniform_standardized = standardize(uniform_data)

normal_normalized = normalize(normal_data)
normal_standardized = standardize(normal_data)

outlier_normalized = normalize(outlier_data)
outlier_standardized = standardize(outlier_data)

# 도표 생성
plt.figure(figsize=(15, 10))

# 균등 분포 데이터 도표
plt.subplot(3, 3, 1)
plt.hist(uniform_data, bins=10, alpha=0.7, label='Original')
plt.title("Uniform Distribution - Original")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.subplot(3, 3, 2)
plt.hist(uniform_normalized, bins=10, alpha=0.7, label='Normalized', color='green')
plt.title("Uniform Distribution - Normalized")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.subplot(3, 3, 3)
plt.hist(uniform_standardized, bins=10, alpha=0.7, label='Standardized', color='orange')
plt.title("Uniform Distribution - Standardized")
plt.xlabel("Value")
plt.ylabel("Frequency")

# 정규 분포 데이터 도표
plt.subplot(3, 3, 4)
plt.hist(normal_data, bins=10, alpha=0.7, label='Original')
```

```
plt.title("Normal Distribution - Original")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.subplot(3, 3, 5)
plt.hist(normal_normalized, bins=10, alpha=0.7, label='Normalized', color='green')
plt.title("Normal Distribution - Normalized")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.subplot(3, 3, 6)
plt.hist(normal_standardized, bins=10, alpha=0.7, label='Standardized', color='orange')
plt.title("Normal Distribution - Standardized")
plt.xlabel("Value")
plt.ylabel("Frequency")

# 이상치를 포함한 데이터 도표
plt.subplot(3, 3, 7)
plt.hist(outlier_data, bins=10, alpha=0.7, label='Original')
plt.title("Data with Outliers - Original")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.subplot(3, 3, 8)
plt.hist(outlier_normalized, bins=10, alpha=0.7, label='Normalized', color='green')
plt.title("Data with Outliers - Normalized")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.subplot(3, 3, 9)
plt.hist(outlier_standardized, bins=10, alpha=0.7, label='Standardized', color='orange')
plt.title("Data with Outliers - Standardized")
plt.xlabel("Value")
plt.ylabel("Frequency")

plt.tight_layout()
plt.show()
```