股東會議資料分析

本筆記本用於分析不同類型股東會議的出席率數據，包含以下分析：

- 基本統計分析
- 常態性檢定
- 變異數同質性檢定
- Mann-Whitney U檢定
- CUSIP重疊分析

1. 基本設置與數據載入

首先導入所需的庫並設置基本類別結構，用於載入和處理數據。 主要功能：

- 導入必要的Python庫
- 設置DataAnalyzer類別
- 實現數據載入功能

In [ ]:
```python
import numpy as np
import pandas as pd
import os
from scipy import stats
from scipy.stats import levene, anderson, kstest
import matplotlib.pyplot as plt
from typing import Dict, List, Set, Tuple

class DataAnalyzer:
    def __init__(self, data_dir: str):
        self.data_dir = data_dir
        self.dataframes = self._load_data()

    def _load_data(self) -> Dict[str, pd.DataFrame]:
        """載入所有CSV檔案
        Returns:
            Dict[str, pd.DataFrame]: 包含各類型會議資料的字典
        """
        try:
            return {
                'Regular': pd.read_csv(os.path.join(self.data_dir, 'regular_events_cleaned_lowest.csv')),
                'Merger': pd.read_csv(os.path.join(self.data_dir, 'condition_merger_agreement_lowest.csv')),
                'Proxy': pd.read_csv(os.path.join(self.data_dir, 'condition_proxy_contest_lowest.csv')),
                'Shareholder': pd.read_csv(os.path.join(self.data_dir, 'condition_shareholder_sponsored_lowest.csv')),
                'Special': pd.read_csv(os.path.join(self.data_dir, 'condition_special_meeting_lowest.csv'))
            }
        except FileNotFoundError as e:
            print(f"錯誤：無法找到檔案 - {e}")
            raise
        except Exception as e:
            print(f"錯誤：資料載入失敗 - {e}")
            raise
```

2. 基本統計分析

計算各類型會議的基本統計指標，包括：

- 樣本數和比例
- 集中趨勢（平均數、中位數）
- 離散程度（標準差、四分位數）
- 分布特徵（偏度、峰度）

In [ ]:
```python
def calculate_basic_stats(self) -> pd.DataFrame:
    """計算基本統計指標"""
    cols = ['樣本數', '佔總樣本比例', '平均數', '中位數', '標準差',
            '最小值', '最大值', '第25百分位', '第75百分位',
            '偏度', '峰度']

    total_samples = sum(len(df) for df in self.dataframes.values())

    stats_data = {}
    for name, df in self.dataframes.items():
        stats_data[name] = [
            len(df),
            f"{(len(df)/total_samples*100):.2f}%",
            *[f"{getattr(df['attendrate'], stat)():.6f}" for stat in
              ['mean', 'median', 'std', 'min', 'max']],
            *[f"{df['attendrate'].quantile(q):.6f}" for q in [0.25, 0.75]],
            f"{df['attendrate'].skew():.6f}",
            f"{df['attendrate'].kurtosis():.6f}"
        ]

    return pd.DataFrame(stats_data, index=cols)
```

3. 數據診斷與統計檢定

進行各種統計檢定，包括：

- 常態性檢定
- 變異數同質性檢定
- Mann-Whitney U檢定
- Dunn's多重比較

In [ ]:
```python
def perform_statistical_tests(self):
    """執行統計檢定"""
    # 常態性檢定
    for name, df in self.dataframes.items():
        self._perform_normality_test(df['attendrate'], name)

    # Levene檢定
    stat, p_value = levene(*[df['attendrate'] for df in self.dataframes.values()])
    print(f"\nLevene's test:\nStatistic: {stat:.4f}\np-value: {p_value:.4f}")

    # Mann-Whitney U檢定
    regular_data = self.dataframes['Regular']['attendrate']
    for name, df in self.dataframes.items():
        if name != 'Regular':
            stat, p = mannwhitneyu(regular_data, df['attendrate'])
            print(f"\nMann-Whitney U test (Regular vs {name}):\n"
                  f"Statistic: {stat:.4f}\np-value: {p:.4f}")
```

4. CUSIP分析

分析不同類型會議之間的CUSIP重疊情況，了解：

- 各類型會議的獨特CUSIP數量
- CUSIP重疊程度
- 重疊比例分析

In [ ]:
```python
def analyze_cusip_overlap(self):
    """分析CUSIP重疊情況"""
    regular_cusips = set(self.dataframes['Regular']['cusip'])
    print(f"Regular Events總CUSIP數: {len(regular_cusips)}")

    for name, df in self.dataframes.items():
        if name != 'Regular':
            condition_cusips = set(df['cusip'])
            overlap = regular_cusips.intersection(condition_cusips)

            print(f"\nRegular vs {name}:")
            print(f"{name}總CUSIP數: {len(condition_cusips)}")
            print(f"重疊CUSIP數: {len(overlap)}")
            print(f"重疊比例: {len(overlap)/len(regular_cusips)*100:.2f}%")
            print(f"獨特CUSIP數: {len(condition_cusips - regular_cusips)}")
```

5. 主程序執行

設置路徑並執行完整分析流程

In [ ]:
```python
def main():
    DIR = r"path/to/your/data/directory"  # 更新為實際數據目錄路徑
    analyzer = DataAnalyzer(DIR)

    print("=== 基本統計分析 ===")
    print(analyzer.calculate_basic_stats())

    print("\n=== 統計檢定 ===")
    analyzer.perform_statistical_tests()

    print("\n=== CUSIP分析 ===")
    analyzer.analyze_cusip_overlap()

if __name__ == "__main__":
    main()
```