

# Part 1: 무역 수출 서프라이즈 정의하기

## 과제 목표

과거 데이터를 기반으로 수출 실적이 예상 대비 얼마나 놀라운지를 측정하는 **surprise 지표**를 만듭니다. 이 지표는 이후 2, 3, 4번 과제의 핵심 입력값으로 사용됩니다.

## Step 1: 기본 성장률 지표 계산하기

### 데이터 이해하기

먼저 `export_value.csv` 파일을 열어보세요. 이 파일은 각 기업(symbol)의 월별 수출액이 기록되어 있습니다. 데이터 구조는 세 개의 컬럼으로 이루어져 있습니다:

- `date`: 수출 데이터 발표일 (매월 말일)
- `symbol`: 기업 코드
- `export_value`: 수출액 (단위: 원)

### YoY (Year-over-Year) 계산

YoY는 전년 동월 대비 성장률입니다. 예를 들어, 2024년 3월의 수출액을 2023년 3월의 수출액과 비교하는 것입니다. 왜 전년 동월과 비교할까요? 수출 데이터에는 강한 계절성이 있기 때문입니다. 연말에는 수출이 급증하고, 연초에는 감소하는 패턴이 반복됩니다. 이런 계절성을 제거하기 위해 정확히 12개월 전 데이터와 비교합니다.

계산 공식은 간단합니다:  $(\text{이번달 수출액} - 12개월 전 수출액) / 12개월 전 수출액$

이를 계산하려면 각 기업별로 데이터를 정렬한 후, 12개월 전 데이터를 가져와야 합니다. pandas의 `groupby`와 `shift` 함수를 사용하면 편리합니다. 기업별로 그룹화한 후, 12개월만큼 `shift`하여 과거 데이터를 가져올 수 있습니다.

### QoQ (Quarter-over-Quarter) 계산

QoQ는 직전 분기 대비 성장률입니다. 분기란 3개월을 의미하며, Q1(1-3월), Q2(4-6월), Q3(7-9월), Q4(10-12월)로 나뉩니다. QoQ를 계산하려면 먼저 각 월이 어느 분기에 속하는지 파악해야 합니다.

계산 과정은 두 단계입니다. 첫째, 각 분기별 평균 수출액을 계산합니다. 예를 들어 2024년 Q1의 평균은  $(1\text{월} + 2\text{월} + 3\text{월}) / 3$ 입니다. 둘째, 이 분기 평균을 직전 분기 평균과 비교합니다. 공식은:  $(\text{이번 분기 평균} - \text{직전 분기 평균}) / \text{직전 분기 평균}$

날짜 컬럼에서 연도와 분기 정보를 추출한 후, 기업별로 분기 평균을 구하세요. 그리고 각 분기의 직전 분기 데이터를 가져와 성장률을 계산합니다.

## MoM (Month-over-Month) 계산

MoM은 전월 대비 성장률로 가장 직관적입니다. 공식은:  $(\text{이번 달 수출액} - \text{지난 달 수출액}) / \text{지난 달 수출액}$

YoY와 마찬가지로 기업별로 그룹화한 후, 1개월 만큼 shift하여 전월 데이터를 가져옵니다. MoM은 가장 최근의 변화를 포착하지만, 계절성과 노이즈에 민감합니다.

## ❷ 중요한 질문: 분모가 0인 경우

이 세 가지 지표를 계산하다 보면 심각한 문제에 직면합니다. 어떤 기업이 특정 달에 수출액이 0원이라면, 그 다음 달의 성장률을 어떻게 계산할까요? 0으로 나누면 무한대가 나옵니다.

**옵션 1: NaN 처리** 가장 단순한 방법은 계산이 불가능한 경우를 결측치(NaN)로 처리하는 것입니다. 이렇게 하면 해당 데이터 포인트는 이후 분석에서 자동으로 제외됩니다. 장점은 단순하고, 거짓 정보를 만들지 않는다는 점입니다. 단점은 데이터가 줄어든다는 점입니다.

**옵션 2: 작은 값(epsilon) 더하기** 분모에 아주 작은 값(예: 0.000001)을 더하는 방법입니다. 이렇게 하면 0으로 나누는 상황을 피할 수 있습니다. 장점은 모든 데이터를 활용할 수 있다는 점입니다. 단점은 인위적으로 수정한 값이라는 점과, epsilon 크기를 어떻게 정할지 애매하다는 점입니다.

**옵션 3: 특별 케이스 플래그** 수출액이 0에서 양수로 바뀌는 경우는 "새로운 수출이 시작됨"을 의미합니다. 이를 별도의 플래그로 표시하고, 성장률 계산에서는 제외하는 방법입니다. 이는 가장 정교하지만 복잡한 방법입니다.

여러분은 세 가지 옵션 중 하나를 선택하거나, 여러 방법을 섞어서 사용할 수 있습니다. 중요한 것은 어떤 방법을 선택했고, 그 이유가 무엇인지 명확히 설명하는 것입니다.

## 기초 통계량 확인

YoY, QoQ, MoM을 모두 계산했다면, 각 지표의 기초 통계량을 확인하세요:

- 평균(mean): 평균적으로 얼마나 성장하는가?
- 표준편차(std): 변동성이 얼마나 큰가?
- 최솟값/최댓값: 극단적인 값이 있는가?

- 결측치 개수: 데이터가 얼마나 손실되었는가?

이 통계량들은 나중에 Z-Score를 계산할 때 핵심이 됩니다.

---

## Step 2: Surprise 개념 이해하기

### Surprise란 무엇인가?

금융에서 "surprise"는 실제 발표된 값이 시장의 예상치와 얼마나 다른지를 의미합니다. 중요한 점은, 단순히 "수출액이 높다"는 것이 놀라운 게 아니라는 점입니다. 만약 어떤 기업이 항상 수출액이 크다면, 그건 이미 모두가 아는 사실입니다. 진짜 놀라운 것은 "예상과 달랐을 때"입니다.

### 실제 사례로 이해하기

회사 A를 생각해봅시다. 이 회사의 최근 6개월 MoM 성장률을 보니 5%, 4%, 6%, 5%, 4%, 5%입니다. 평균적으로 매달 약 5% 성장합니다. 그런데 이번 달 갑자기 15% 성장했습니다. 이게 바로 "positive surprise"입니다.

반대로, 항상 월 5%씩 성장하던 회사가 이번 달 -3% 성장(즉, 감소)했다면, 이는 "negative surprise"입니다.

### 왜 중요한가?

주식 시장은 효율적입니다. 투자자들은 기업의 평균적인 성장률을 이미 알고 있고, 이를 주가에 반영합니다. 예를 들어, "이 회사는 매달 5% 성장한다"는 정보는 이미 주가에 포함되어 있습니다.

하지만 예상과 다른 결과가 나오면 주가가 움직입니다. 15% 성장이라는 "깜짝" 소식은 새로운 정보이므로 주가가 상승합니다. 반대로 -3% 성장이라는 "실망" 소식은 주가를 하락시킵니다.

우리의 목표는 바로 이 "예상과의 차이"를 수치화하는 것입니다. 단순히 YoY가 높은 회사를 찾는 게 아니라, "평소 패턴 대비 얼마나 이례적인가"를 측정해야 합니다.

---

## Step 3: Z-Score를 이용한 표준화 (10점)

### Z-Score의 기본 개념

Z-Score는 통계학에서 가장 많이 사용되는 표준화 기법입니다. 핵심 아이디어는 간단합니다: "이 값이 평균으로부터 표준편차 단위로 얼마나 떨어져 있는가?"

공식은:  $z = (x - \mu) / \sigma$

- $x$ : 관측값 (예: 이번 달 YoY)
- $\mu$ : 평균 (예: 과거 YoY의 평균)
- $\sigma$ : 표준편차 (예: 과거 YoY의 표준편차)

## Z-Score의 해석

Z-Score = 0이면 정확히 평균입니다. Z-Score = 1이면 평균보다 1 표준편차 위에 있다는 뜻입니다. 정규분포를 가정할 때, 전체 데이터의 약 84%가 Z-Score 1 이하에 있습니다. 즉, Z-Score 1은 상위 16%에 해당합니다.

Z-Score = 2라면 상위 2.5%에 해당하는 매우 이례적인 값입니다. Z-Score = 3이면 상위 0.1%로, 극히 드문 사건입니다.

음수 Z-Score도 동일하게 해석합니다. Z-Score = -2는 평균보다 2 표준편차 아래, 즉 하위 2.5%에 해당합니다.

## 왜 Z-Score를 사용하는가?

서로 다른 기업을 비교하고 싶다고 가정해봅시다. A기업의 YoY는 평균 10%, 표준편차 5%입니다. B기업의 YoY는 평균 50%, 표준편차 30%입니다. A기업이 이번 달 20% 성장했고, B기업이 80% 성장했다면 어느 쪽이 더 놀라운 결과일까요?

절댓값으로 보면 B기업이 더 크게 성장했지만, Z-Score로 계산하면:

- A기업:  $(20 - 10) / 5 = 2$
- B기업:  $(80 - 50) / 30 = 1$

A기업의 성장이 더 이례적입니다! Z-Score는 각 기업의 "평소 패턴"을 고려하여 표준화하므로, 서로 다른 기업을 공정하게 비교할 수 있습니다.

## 전체 기간 Z-Score 계산

먼저 간단한 버전부터 시작합니다. 각 기업의 전체 기간 YoY 데이터를 사용하여 평균과 표준편차를 구하고, 이를 바탕으로 Z-Score를 계산합니다.

기업별로 그룹화한 후, YoY 컬럼의 평균과 표준편차를 계산합니다. 그리고 각 데이터 포인트에서 평균을 빼고 표준편차로 나눕니다. pandas의 `transform` 함수를 사용하면 그룹별 계산이 편리합니다.

## 분포 시각화

Z-Score를 계산했다면 히스토그램을 그려보세요. 정규분포를 따른다면 종 모양의 그래프가 나타날 것입니다. X축에  $\pm 2$  표준편차 선을 그려서 이상치(outlier)의 비율을 시작적으로 확인하세요.

실제로  $|Z\text{-Score}| > 2$ 인 데이터 포인트가 전체의 몇 퍼센트인지 계산해보세요. 이론적으로는 약 5%여야 하지만, 실제 금융 데이터는 종종 "fat tail"(극단값이 이론보다 많이 나타남)을 보입니다.

## 이상치 분석

Z-Score가 가장 큰 10개 사례와 가장 작은 10개 사례를 찾아보세요. 이들은 가장 극단적인 positive/negative surprise입니다. 해당 날짜와 기업을 확인하고, 실제로 무슨 일이 있었는지 조사해보면 흥미로운 인사이트를 얻을 수 있습니다.

## Step 4: Rolling Window 선택하기

### 문제 인식

전체 기간 Z-Score에는 치명적인 문제가 있습니다. 미래 정보를 사용한다는 점입니다. 예를 들어, 2022년 6월의 Z-Score를 계산할 때 2024년 데이터까지 사용했다면, 이는 현실에서 불가능합니다. 2022년 6월 시점에서는 2024년 데이터를 알 수 없기 때문입니다.

이를 "Forward-Looking Bias" 또는 "Look-Ahead Bias"라고 합니다. 이는 백테스트의 치명적 오류이며, 실제 수익률과 괴리를 만듭니다.

해결책은 Rolling Window입니다. 특정 시점의 Z-Score를 계산할 때, 그 시점까지의 과거 데이터만 사용하는 것입니다.

### 데이터 손실 Trade-off

Rolling Window를 사용하면 데이터가 손실됩니다. 24개월 window를 사용한다면, 처음 24개월은 Z-Score를 계산할 수 없습니다. 왜냐하면 충분한 과거 데이터가 없기 때문입니다.

현재 60개월 데이터가 있다면:

- 12개월 window: 48개월 사용 가능
- 24개월 window: 36개월 사용 가능
- 36개월 window: 24개월 사용 가능

여기서 Trade-off가 발생합니다. 짧은 window는 최근 트렌드를 빠르게 반영하지만 추정이 불안정합니다. 긴 window는 안정적이지만 구조적 변화를 느리게 반영하고 데이터를 많이 잃습니다.

## 계절성 고려

수출 데이터는 강한 계절성을 갖습니다. 연말에는 수출이 급증하고, 명절 기간에는 감소합니다. 이런 계절성을 제대로 포착하려면 최소한 12개월(1년)의 데이터가 필요합니다.

더 나아가, 계절성이 해마다 약간씩 다를 수 있으므로 2년(24개월) 데이터를 보는 것이 안정적입니다. 월별 평균 수출액을 계산하여 계절 패턴을 시각화해보세요.

## 권장 사항: 24개월 Window

대부분의 경우 24개월 rolling window를 권장합니다. 이유는:

1. 계절성을 2회 포함하여 패턴을 안정적으로 추정
2. 최근 2년 트렌드를 반영하여 현실성 유지
3. 36개월의 사용 가능한 데이터 포인트 확보
4. YoY는 이미 12개월 차이를 보므로, 24개월이면 최소 2개의 YoY 비교 가능

하지만 이는 절대적 규칙이 아닙니다. 여러 window size를 실험해보고, 결과를 비교하여 가장 적합한 것을 선택하세요. 선택의 근거를 명확히 제시하는 것이 중요합니다.

---

## Step 5: Rolling Z-Score 계산하기

### Rolling의 의미

Rolling Z-Score는 "움직이는 창문"처럼 작동합니다. 2023년 6월의 Z-Score를 계산한다면, 2021년 7월부터 2023년 6월까지의 24개월 데이터만 사용합니다. 2023년 7월의 Z-Score를 계산한다면, 창문이 한 달 이동하여 2021년 8월부터 2023년 7월까지 사용합니다.

이렇게 하면 각 시점에서 "그때 사용 가능했던 정보"만 활용하게 됩니다.

### 계산 방법

Rolling Z-Score 계산은 다음 단계를 반복합니다:

1. 특정 날짜  $t$ 를 선택
2.  $t$ 로부터 과거 24개월 데이터를 추출 ( $t-24 \sim t$ )

3. 이 24개월 데이터의 평균( $\mu$ )과 표준편차( $\sigma$ )를 계산
4. t 시점의 값에서  $\mu$ 를 빼고  $\sigma$ 로 나눔
5. 다음 날짜로 이동하여 반복

pandas의 `rolling` 함수를 사용하면 이를 효율적으로 구현할 수 있습니다. 기업별로 그룹화한 후, `rolling mean`과 `rolling std`를 계산하고, 이를 사용하여 Z-Score를 구합니다.

## YoY vs QoQ vs MoM의 Window 선택

YoY, QoQ, MoM에 동일한 window를 사용해야 할까요? 반드시 그렇지는 않습니다.

YoY는 이미 12개월 차이를 측정하므로, 장기 트렌드를 반영합니다. 따라서 24~36개월 window가 적합합니다.

MoM은 단기 변동을 측정하므로, 12~18개월 window가 더 적절할 수 있습니다. 너무 긴 window는 MoM의 장점(빠른 반응)을 상쇄합니다.

QoQ는 중간 수준이므로 18~24개월이 적당합니다.

여러분이 선택한 window와 그 이유를 명확히 문서화하세요.

## 검증: Forward-Looking Bias 체크

Rolling Z-Score가 제대로 계산되었는지 검증해야 합니다. 특정 날짜의 Z-Score를 선택하고, 수동으로 재계산해보세요. 그 날짜까지의 과거 24개월 데이터만 사용하여 평균과 표준편차를 구하고, Z-Score를 계산합니다. 이 값이 프로그램이 계산한 값과 일치하는지 확인하세요.

만약 차이가 있다면, 미래 데이터를 사용했을 가능성이 높습니다.

## 결측치 처리

Rolling Z-Score 계산 중 결측치가 발생할 수 있습니다. 주요 원인은:

1. 초기 window 기간 (처음 24개월)
2. 원본 데이터(YoY, QoQ, MoM)의 결측치
3. Window 내 데이터가 너무 적어 표준편차를 계산할 수 없는 경우

각 경우를 어떻게 처리할지 결정하세요. 일반적으로 window 내 최소 데이터 포인트 수를 설정합니다(예: 최소 12개월). 이보다 적으면 Z-Score를 계산하지 않습니다.

## 시계열 시각화

Rolling Z-Score의 시계열 그래프를 그려보세요. X축은 날짜, Y축은 Z-Score입니다.  $\pm 2$  표준편차 선을 함께 그려서 이상치를 표시합니다.

특정 기업 몇 개를 선택하여 개별 그래프를 그려보세요. Z-Score가 ±2를 넘어서는 시점이 언제인지, 그 시점에 실제로 어떤 일이 있었는지 확인해보면 흥미롭습니다.

## 최종 데이터 저장

모든 계산이 끝나면 결과를 CSV 파일로 저장하세요. 파일에는 다음 컬럼이 포함되어야 합니다:

- date: 날짜
- symbol: 기업 코드
- export\_value: 원본 수출액
- yoy, qoq, mom: 성장을
- rolling\_zscore\_yoy, rolling\_zscore\_qoq, rolling\_zscore\_mom: Rolling Z-Score

이 파일이 Part 2, 3, 4의 입력 데이터가 됩니다. 파일명은 `export_with_surprise.csv`로 저장하는 것을 권장합니다.

- 파일을 다시 열어서 데이터가 정상인지 확인했는가?

---

이 단계에서 계산한 `rolling_zscore_yoy`, `rolling_zscore_qoq`, `rolling_zscore_mom`이 바로 우리가 정의한 "Surprise" 지표입니다. 이는 단순히 "수출액이 크다"가 아니라 "평소 패턴 대비 얼마나 이례적인가"를 측정합니다.

Surprise 지표는 금융에서 매우 중요한 개념입니다. 애널리스트들은 매 분기 기업의 실적을 예상하고, 실제 발표된 실적과 비교합니다. "실적 서프라이즈"가 크면 주가가 크게 움직입니다. 우리는 동일한 개념을 수출 데이터에 적용한 것입니다.