

금융 통계

1. 대푯값

중앙값

중앙 값은 자료 크기 나열 후 가운데 위치한 값 (개수가 짝수면 2개 사이의 평균)

절사평균

이상치를 일정 비율 제거 후 구한 평균

```
stats.trim_mean(data, 0.1) # 양쪽 극단 치 10% 제거
```

단순 이동 평균

단순히 주어진 기간 동안의 평균을 구한다.

지수 이동 평균

최근 데이터를 더 중요하게 생각 해 최신의 데이터에 가중치를 더 높게 부여

```
ma5 = chart['KOSPI'].rolling(5).mean() // 5일 이동 평균
ma25 = chart['KOSPI'].rolling(25).mean() // 25일 이동 평균

ewma12 = chart['KOSPI'].ewm(span=12).mean()
```

페어 트레이딩

두 개의 상호관련성 높은 주식을 쌍으로 선택 → 둘 간의 가격차이를 이용해 수익을 추구하는 투자 전략

2. 분산과 표준편차

- 데이터의 흩어진 정도 → 금융에서는 리스크를 측정하는 지표
- 분산 = 표준편차²

활용 방법

- 시장 변동성 측정
- VIX - S&P 500 지수 옵션에 기반한 변동성을 측정하는 방식
- 표준편차와 리스크/투자성향

변동 계수

여러 자료의 표본 평균이 같다? → 각 자료의 표준편차를 비교해 자료의 퍼져있는 정도를 판단 가능

- 평균치 한 단위당 변동이 얼마나 되는지 나타낸다.

주가 변동성 측정 방식

- 하방 변동성: 기대하는 것보다 낮은 수익률에 대한 변동성 측정
- 중앙값 절대 편차(MAD): 평균값으로 구한 표준편차 → 극단치의 영향을 받음 → 데이터의 중앙값과 각 값의 차이의 절대값들의 중앙값으로 데이터의 흩어짐을 측정
 - MAD 식

$$MAD = \text{median}(|X - \text{median}(X)|)$$

3. 왜도와 첨도

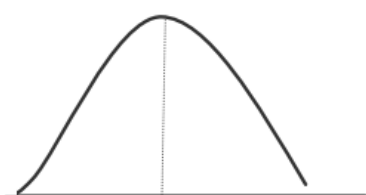
왜도

자료의 분포가 기울어진 방향과 정도를 나타낸다.

$$S_k = \frac{\sum (x - \bar{x})^3 / n}{s^3}$$

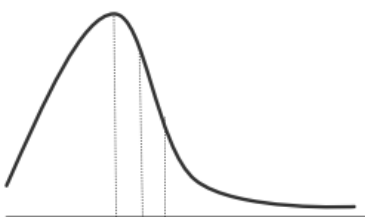
$$S_k = \frac{[E(X - \mu)^3]}{\sigma^3}$$

최빈값 = 중앙값 = 평균



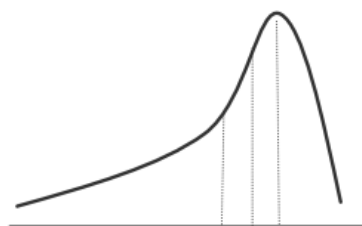
(a) 대칭 분포

최빈값 < 중앙값 < 평균



(b) 양(+)의 왜도 분포

평균 < 중앙값 < 최빈값



(c) 음(-)의 왜도 분포

첨도

자료의 분포가 뾰족한지, 꼬리가 두터운지 측정하는 정도

$$S_k = \frac{\sum (x - \bar{x})^4 / n}{s^4}$$

$$S_k = \frac{[E(X - \mu)^4]}{\sigma^4}$$

4. 사분위수와 백분위수

백분위수는 자료를 100등분하는 값이다.

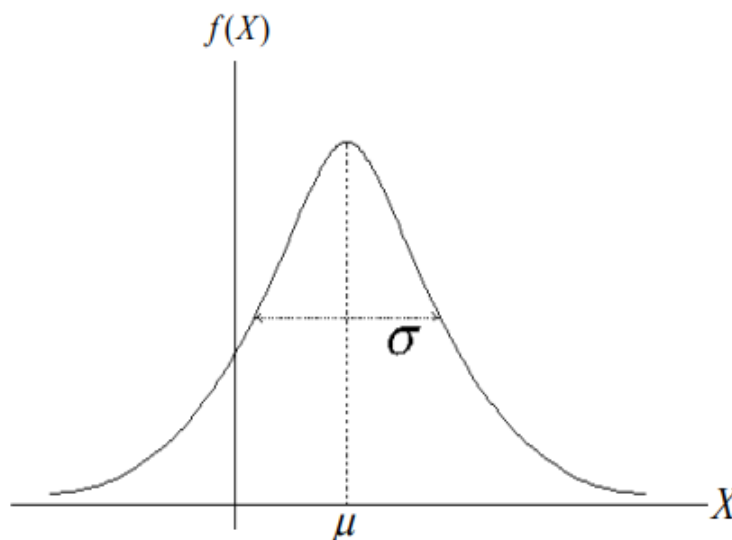
4분위수는 자료를 4등분 하여 Q1, Q2, Q3, Q4로 나눈다.

제 100백분위수 = Q4, 제 75백분위수 = Q3, 제 50백분위수 = Q2, 제 25백분위수 = Q1

- 사분위수 범위(IQR) = Q3 - Q1
- 사분위수 최소치 Q1 - 1.5*IQR
- 사분위수 최대치 Q3 + 1.5*IQR

5. 정규 분포

정규 분포: N(평균, 표준편차^2)



Z-score

$$z = \frac{x - \text{평균}}{\text{표준편차}} = \frac{x - \bar{x}}{s}$$

[표준정규분포] $z = \frac{X - \mu}{\sigma}$

중심 극한 정리

- 모집단의 분포와 상관 없이, 일단 표본의 크기가 충분하면 표본 평균들의 분포가 모집단의 모수를 기반으로한 정규 분포를 이룬다.
- 표본 평균들이 이루는 표본 분포와 모집단 간의 관계를 증명한다.
수집한 표본의 통계량으로 모집단의 모수를 추정할 수 있는 수학적 근거를 마련한다.

VaR (Value of Risk)

주어진 신뢰 수준 하에서 목표 기간 동안 발생할 수 있는 최대 손실 금액

예제

평균 60.36, 표준편차 18.6052 일 때 Z 분포 상위 10%에 해당하는 시험 점수는?

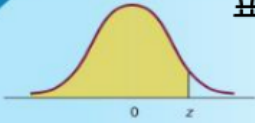
평균 + (Z + 표준편차) = X = 60.36 + 1.2816 × 18.6052 ≈ 84.20

여기서 Z는 Z 분포에서 상위 10%에 해당하는 값이다.

▼ Z-분포

상위 10% (하위 90%)에 가장 가까운 Z 값은 1.2 + 0.08 = 1.28이다.

표준정규분포표(standard normal distribution table)



$P(-\infty < Z < z)$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

6. 공분산

두 확률 변수의 변화 양상을 측정한다.

$$Cov(X, Y) = E[X * Y] - E(X)E(Y)$$

만약 두 변수 사이가 독립 → 선형관계가 아니면 공분산은 0이 된다.

7. 상관 계수 (피어슨)

공분산은 두 확률 변수 X와 Y의 선형 관계를 나타낸다.

즉, 공분산은 X와 Y의 측정 단위에 따라 그 값이 달라짐

단위에 관계 없이 X와 Y의 밀접한 정도를 측정하기 위해 상관계수를 사용한다.

- +1 완전한 정의 관계, -1 완전한 부의 관계, 0 아무런 관계가 없음

스피어맨 상관계수

순위가 매겨지는 변수간의 Pearson 상관계수로 정의된다.

8. 회귀 분석

두 개 이상의 변수들 사이의 관련성을 통해 한 변수의 변화에 따른 다른 변수의 변화를 예측

영향을 주는걸 독립변수(x), 영향 받는 변수를 종속 변수 (y)

- 기본 가정
 - 분산은 동일하다. (동 분산성)
 - 변수들은 서로 독립이다. (독립성)

$$Y_i = \alpha + \beta X_i + e_i (i=1, 2, \dots, n)$$

회귀직선 기울기 = B

B가 0이면 x와 y 사이에 선형 관계 존재 X
B가 양수면 x와 y 사이에 양의 선형 관계 O

결정 계수

총변동 TSS = RSS(설명된 변동) + ESS(설명 안되는 변동 = 에러)

결정계수 $R^2 = RSS / TSS$

다중 회귀 분석

종속 변수 Y에 영향을 미치는 독립 변수가 K개(여러개)있다.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

여기서 각 β 는 해당 독립 변수를 제외한 나머지 (K-1)개의 다른 독립 변수를 고정시켰을 때 해당 독립 변수 변화에 따른 종속 변수의 평균 변화량이다. (당연한 말)

다중 공선성 문제

VIF가 10이 넘으면 다중공선성 의심 (연관있는 변수끼리 VIF가 높음)

다중 공선성이란 독립 변수들 간의 완전한 선형 종속의 관계를 칭한다.

다중공선성이 존재 시 회귀 분석 독립 변수는 상호 독립이라는 가정에 위배된다.

회귀계수에 대한 추론

결정 계수 값(R-squared)

X가 Y를 얼마나 잘 설명하는 지, 1에 가까울 수록 X가 Y를 잘 설명한다는 뜻

Prob(p값)

회귀 분석에서 사용한 X와 Y가 통계적으로 유의미한 영향을 미치는지 여부

p값과 유의수준(알파) 비교

유의수준은 오류를 허용할 최대 확률 (보통 0.05, 5%) → 귀무가설 기각 여부를 판단

p값이 0.05보다 작을 시 귀무가설을 기각하고 대립가설을 채택한다는 뜻.

p값이 유의수준(0.05)보다 작으면 X가 Y에 유의미한 영향을 미친다고 생각

검정은 t검정 f검정이 있음

결정계수값이 높아야한다. (R-스퀘어)

Prob 는 내가 틀릴 확률, 잘못될 확률로써 이해하자.

t값과 p값을 통해서 유의 검증을 진행할 수 있다.

p값이 유의수준(95%, 0.05)보다 작으므로 통계적으로 유의하다는 것을 알 수 있다.

단순회귀분석 : 판매량 = f (광고비)

OLS Regression Results

Dep. Variable: sales

R-squared: 0.925

Model: OLS

Adj. R-squared: 0.916

Method: Least Squares

F-statistic: 98.88

Date: Sat, 01 Jun 2024

Prob (F-statistic): 8.85e-06

Time: 21:21:27

Log-Likelihood: -68.092

No. Observations: 10

AIC: 140.2

Df Residuals: 8

BIC: 140.8

Df Model: 1

Covariance Type: nonrobust

회귀계수의 표준오차

검정통계량(>|2.0|)

P-value (양측검정)

유의수준=0.05

	coef	std err	t	P> t	[0.025	0.975]
const	41.3956	119.559	0.346	0.738	-234.309	317.100
ad_cost	9.2927	0.935	9.944	0.000	7.138	11.448

Omnibus: 5.027

Prob(Omnibus): 0.081

Skew: -0.610

Kurtosis: 4.315

Durbin-Watson: 오차항 자기상관 ≈ 2

Jarque-Bera (JB): 정규분포 여부

Prob(JB): 0.511

Cond. No. 197.

$dw = 2(1 - \rho)$

귀무가설: 잔차항은 정규분포를 따른다.

prod: 내가 잘못될 확률, 값이 작다 → 틀릴 확률이 적다.