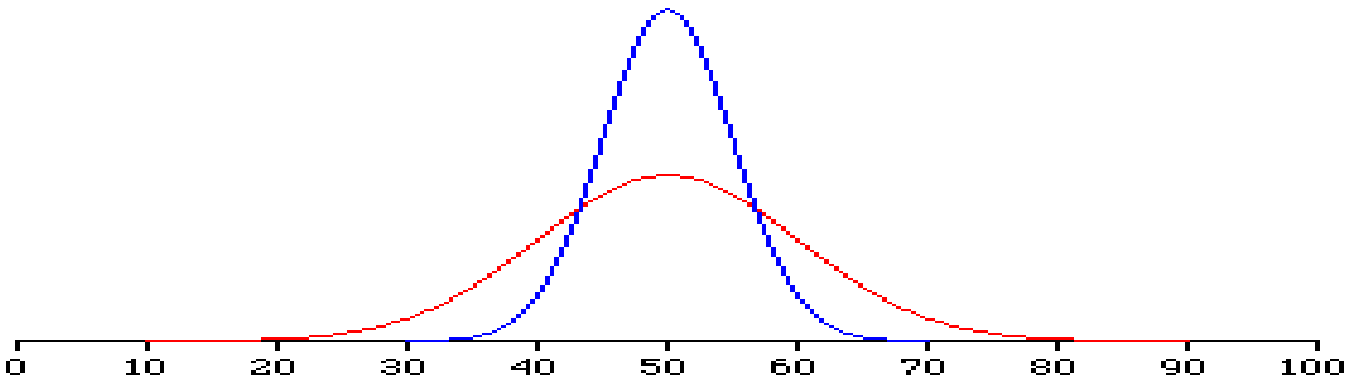
자료분석방법II:
수치를 이용한 기술적 통계분석

변동성(Variability)의 측정

- 중심위치의 측정은 자료의 분포(distribution), 즉 변수값들이 평균(중심위치)를 중심으로 얼마만큼 산포되어 있는지, 에 대해서는 어떤 특성이나 정보를 제공하지 못한다.



- 비록 두 자료가 같은 평균값(50)을 가지고 있으나 빨간색의 자료가 파란색의 자료 보다 변동성이 크다

변동성의 측정: 범위(Range)

- **범위(*range*)는 변동성을 측정하는데 있어서 가장 간단하고 단순한 방법이다**

범위 = 가장 큰 변수값 – 가장 작은 변수값

Ex).

Data: {4, 4, 4, 4, 50} 범위 = (50-4) = 46

Data: {4, 8, 15, 24, 39, 50} 범위 = (50-4) = 46

→ 두 자료의 경우 같은 범위의 값을 가지나 두 자료의 분포는 매우 다른 형태를 띠고 있다

변동성의 측정: 범위(Range)

-장점: 아주 쉽게 산출 할 수 있다.

-단점: 두 개(가장 큰 수와 가장 작은 수)의 변수값 사이에 있는 다른 변수 값들의 변동성 정도에 대한 정보를 제공하지 못한다

- 따라서 두개의 변수값에 의한 변동성이 아닌 모든 변수값들에 대한 변동성 측정이 필요하다…

변동성의 측정: 평균편차(Mean Deviation)

MEAN DEVIATION

$$MD = \frac{\Sigma|X - \overline{X}|}{n}$$

Ex) 춘천의 Starbucks 커피점에서 하루에 판매되는 cappuccinos 수를 조사한 결과 지난 5일간 20, 40, 50, 60, and 80 이었다고 하자. 이 표본자료에서 cappuccinos 수의 평균편차값을 구하면?

→ 평균($\overline{x}$) = (20+40+50+60+80)/5 = 50

Number of Cappuccinos Sold Daily	$(X - \overline{X})$	Absolute Deviation
20	(20 − 50) = −30	30
40	(40 − 50) = −10	10
50	(50 − 50) = 0	0
60	(60 − 50) = 10	10
80	(80 − 50) = 30	30
		Total 80

$$MD = \frac{\Sigma|X - \overline{X}|}{n} = \frac{80}{5} = 16$$

변동성의 측정: 분산(Variance)과 표준편차 (standard Deviation)

- 분산과 표준편차는 평균과 함께 매우 중요한 통계지수 이다

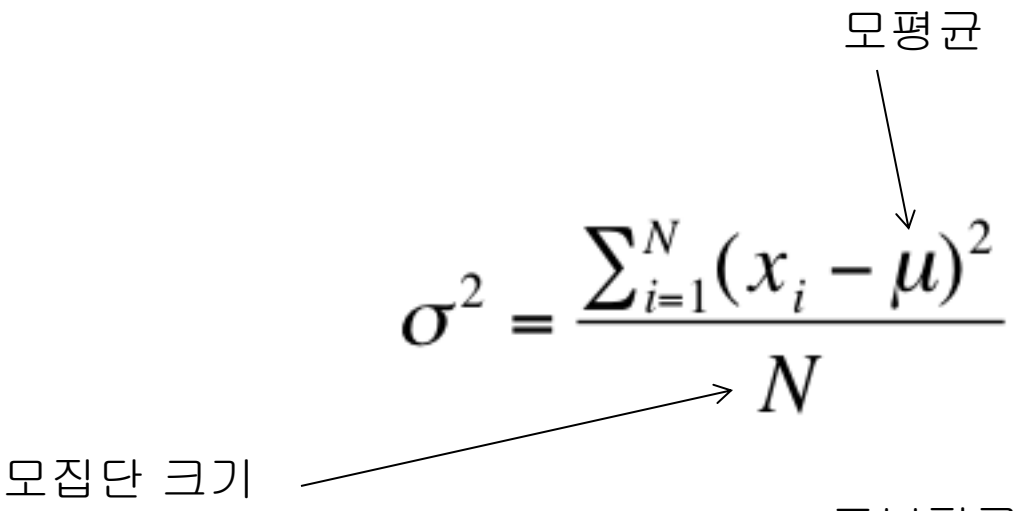
- 변동성 측정이외에 통계적 추정과정(statistical inference procedures)에서 매우 중요한 역할을 수행한다

- 모집단 분산(모분산:Population variance)은 σ^2 로 표기(Greek letter “sigma” squared)

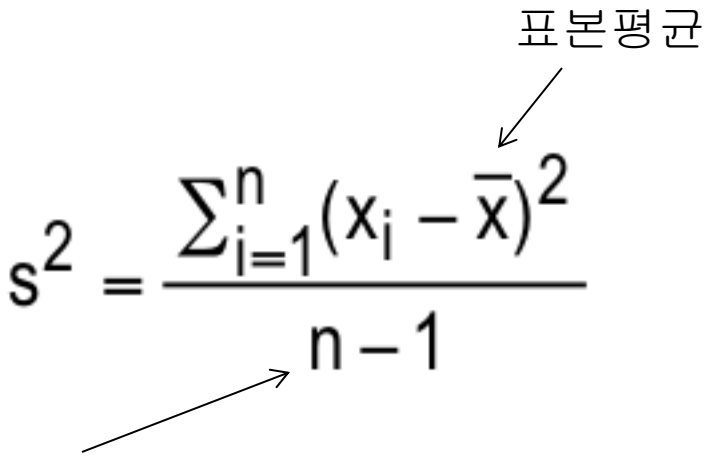
- 표본분산(Sample variance)은 s^2 로 표기 (“S” squared)

분산(Variance)값의 산출

모분산 값은



표본분산 값은



주의: 분모값은 표본크기 **(n)-1**

분산(Variance)값의 산출

- 분산 값의 산출식에서 보듯이 분산값을 구하기 위해 먼저 평균값을 산출해야 한다

- **평균값 없이 분산값을 구하는 방법(표본분산)**

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

예시:표본평균과 표본분산

표본자료(X)가 다음과 같이 6개의 값으로 구성되었다고 가정하자.

X = {17, 15, 23, 7, 9, 13}

- 표본평균과 표본분산값을 구하시오

표본평균과 표본분산

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{17+15+23+7+9+13}{6} = \frac{84}{6} = 14.$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{6-1}\left[(17-14)^2 + (15-14)^2 + \ldots(13-14)^2\right] = 33.2$$

Sample Variance (shortcut method)

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right] = \frac{1}{6-1}\left[\left(17^2+15^2+\ldots+13^2\right) - \frac{(17+15+\ldots+13)^2}{6}\right] = 33.2$$

표준편차(Standard Deviation)

- 표준편차는 분산값에 root를 취한 값이다:

모집단의 표준편차(Population standard deviation)

$$\sigma = \sqrt{\sigma^2}$$

표본의 표준편차(Sample standard deviation)

$$s = \sqrt{s^2}$$

표준편차의 해석

Ex) 어느 골프제조업자가 새 골프클럽을 만들어 현재의 골프클럽과 비교하기 위해 거리를 실험한 결과 다음과 같은 표를 얻었다

Current 7-iron	
Mean	150.55
Standard Error	0.67
Median	151
Mode	150
Standard Deviation	5.79
Sample Variance	33.55
Kurtosis	0.13
Skewness	-0.43
Range	28
Minimum	134
Maximum	162
Sum	11291
Count	75

New 7-iron	
Mean	150.15
Standard Error	0.36
Median	150
Mode	149
Standard Deviation	3.09
Sample Variance	9.56
Kurtosis	-0.89
Skewness	0.18
Range	12
Minimum	144
Maximum	156
Sum	11261
Count	75

새 클럽의 표준편차가 작게 나타남에 따라 새클럽이 현재의 클럽보다 거리변동에 있어서 보다 **일관성**있음을 알수있다

표준편차의 응용: 경험적 법칙(Empirical Rule)

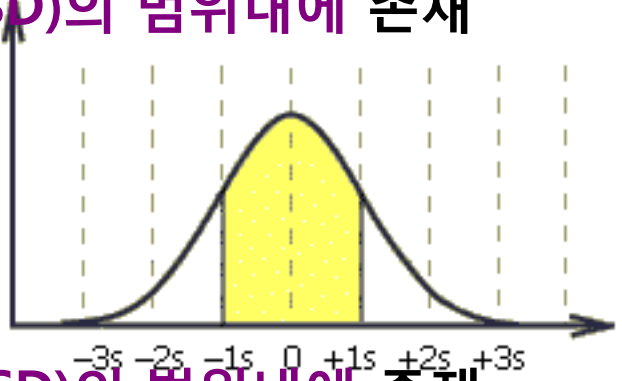
- 표준편차는 서로다른 분포형태를 가진 자료들의 변동성을 비교하는데 이용되며, 아울러 자료의 분포특성을 파악하는데 사용된다

- 만일 자료의 분포형태를 나타내는 히스토그램이 대칭적 형태(종모양의 형태:bell shape)일 경우 , 경험적 법칙(*Empirical Rule*)을 적용할 수 있다

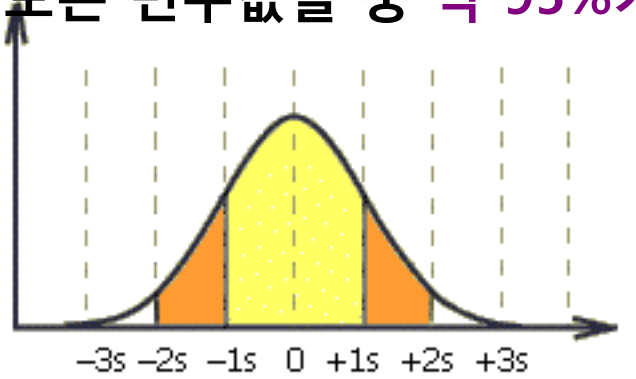
1) 모든 변수값들 중 약 65%가 평균±표준편차의 범위내에 존재한다
2) 모든 변수값들 중 약 95%가 평균±(2*표준편차)의 범위내에 존재한다
3) 모든 변수값들 중 약 99.7%가 평균±(3*표준편차)의 범위내에 존재한다

경험적 법칙(Empirical Rule)

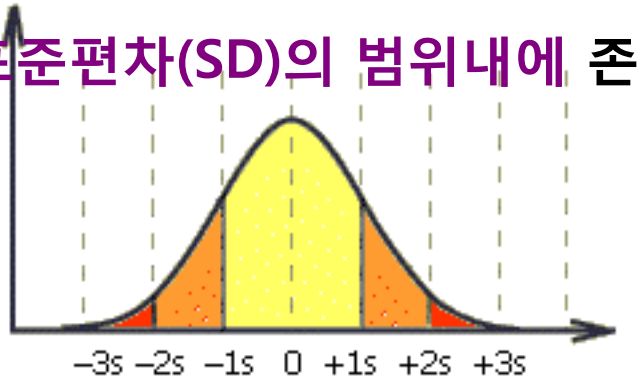
1) 모든 변수값들 중 약 65%가 평균(0)±표준편차(SD)의 범위내에 존재



2) 모든 변수값들 중 약 95%가 평균(0)±2*표준편차(SD)의 범위내에 존재



3) 모든 변수값들 중 약 99.7%가 평균(0)±3*표준편차(SD)의 범위내에 존재



체비세프 이론(Chebysheff's Theorem)

- 주어진 자료의 분포가 비대칭적 형태인 경우 (대칭적 종 모양의 형태가 아니 경우)에 ***Chebysheff's Theorem***,을 적용할 수 있다

- **주어진 자료의 모든 변수값들의 100*[$1-\frac{1}{k^2}$]%가 적어도 [mean±(k*SD)] 범위내에 존재한다**

→ 만일 k=2 일 경우, 체비세프 이론에 의하면, 모든 변수값들의 75%[1-(1/2^2)]가 **[mean±(2*SD)]**범위내에 존재한다