



부산대학교

PUSAN NATIONAL UNIVERSITY

Derivatives and Probability

Derivatives and Probability

Local Change and Uncertainty in Model Fitting

BX3600087-140 / DS3600087-140 / AB3600087-140

기계학습개론

2026학년도 2학기

오용경

데이터사이언스전공

yongkyung.oh@pusan.ac.kr

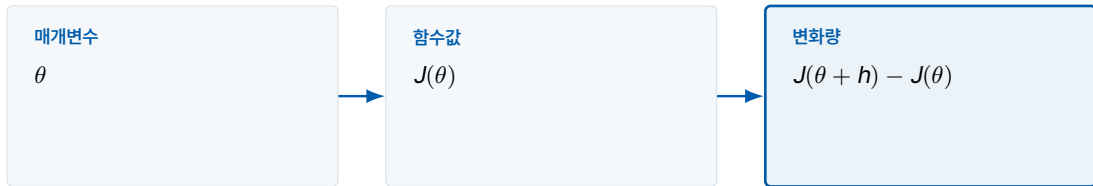
Section 1 • Derivatives and Probability

Local Change of a Function

매개변수의 작은 변화가 함수값에 미치는 영향

매개변수의 작은 변화

매개변수 θ 를 h 만큼 바꿀 때 함수값이 얼마나 달라지는지 확인한다



도함수(derivative)는 매개변수 변화에 대한 함수값의 국소 변화율이다

함수값의 변화

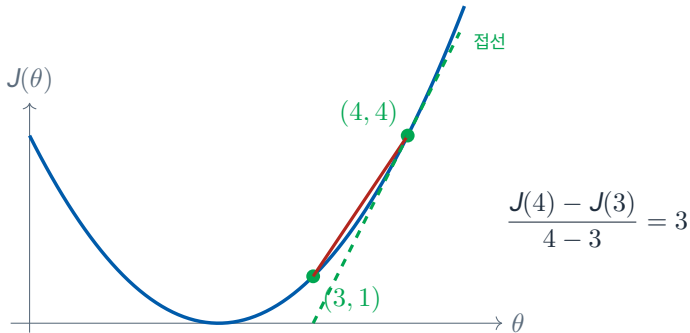
실수 하나를 입력받아 실수 하나를 출력하는 함수부터 확인한다

θ	$J(\theta) = (\theta - 2)^2$	$\theta - 2$
0	4	-2
2	0	0
4	4	2

$\theta = 2$ 에서 함수값이 가장 작고, 양쪽으로 멀어질수록 증가한다

두 점을 잇는 직선의 기울기

$\theta = 3$ 과 $\theta = 4$ 사이에서 함수값이 변한 비율을 계산한다



두 점을 잇는 직선의 기울기(slope)는 두 점 사이의 평균변화율이다. 접선은 $\theta = 4$ 에서 곡선에 닿는 접선(기울기 4)이다

Source: Zhang et al. (2023), Derivatives and Differentiation.

간격을 줄이면

$\theta = 4$ 를 기준으로 다른 점을 가까이 옮기며 평균변화율을 비교한다

h	$J(4 + h)$	$\frac{J(4 + h) - J(4)}{h}$	평균변화율
-1	$J(3) = 1$	$(1 - 4)/(-1)$	3
-0.5	$J(3.5) = 2.25$	$(2.25 - 4)/(-0.5)$	3.5
-0.1	$J(3.9) = 3.61$	$(3.61 - 4)/(-0.1)$	3.9
0.1	$J(4.1) = 4.41$	$(4.41 - 4)/0.1$	4.1

h 가 양쪽에서 0에 가까워질수록 평균변화율은 4에 가까워진다

도함수의 정의

간격 h 를 0으로 줄인 극한이 한 점에서의 변화율이다

$$J'(\theta) = \lim_{h \rightarrow 0} \frac{J(\theta + h) - J(\theta)}{h}$$

기호	뜻
h	매개변수에 더하는 작은 변화
$J(\theta + h) - J(\theta)$	함수값의 변화
$J'(\theta)$	θ 에서의 국소 변화율. 두 점을 잇는 직선이 $h \rightarrow 0$ 에서 가까워지는 접선의 기울기

Source: Zhang et al. (2023), Derivatives and Differentiation.

도함수의 부호와 변화 방향

작은 h 에서 $J(\theta + h) - J(\theta) \approx J'(\theta)h$ (1차 근사)로 증가와 감소를 판단한다

$$J'(\theta) < 0$$

θ 를 늘리면 J 가 감소

$$J'(\theta) = 0$$

1차 근사에서 변화가 없음

$$J'(\theta) > 0$$

θ 를 늘리면 J 가 증가

부호는 함수값이 늘어나는지 줄어드는지를, 절댓값은 같은 작은 변화에 함수값이 얼마나 크게 변하는지를 나타낸다

기본 미분 공식

규칙	함수	도함수	적용하는 항
상수	c	0	고정된 항
거듭제곱	θ^n	$n\theta^{n-1}$	거듭제곱 항
합	$f + g$	$f' + g'$	여러 항의 합
상수배	$cf(\theta)$	$cf'(\theta)$	계수가 붙은 항

합으로 이루어진 식은 각 항에 미분 공식을 적용하여 계산한다

Source: Zhang et al. (2023), Derivatives and Differentiation.

이차함수를 미분한다

식을 펼친 뒤 거듭제곱과 합의 미분 공식을 적용한다

$$J(\theta) = (\theta - 2)^2 = \theta^2 - 4\theta + 4$$

$$J'(\theta) = 2\theta - 4 = 2(\theta - 2)$$

$$J'(4) = 2(4 - 2) = 4$$

$$J'(1) = 2(1 - 2) = -2$$

$\theta = 4$ 에서는 θ 를 조금 줄이고, $\theta = 1$ 에서는 조금 늘리면 함수값이 감소한다

Checkpoint

Local Change of a Function

- $J'(0), J'(2), J'(4)$ 의 값은 각각 얼마인가
- 각 점에서 θ 를 어느 쪽으로 움직여야 J 가 감소하는가
- $J'(2) = 0$ 은 1차 근사에서 무엇을 뜻하는가

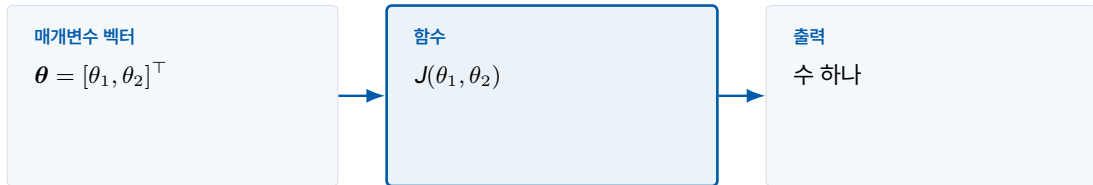
Section 2 • Derivatives and Probability

Multivariable Gradients

편미분을 모아 그래디언트를 구성

매개변수가 둘인 함수

J 는 벡터(vector)를 입력받아 수 하나를 출력한다



매개변수가 여러 개이면 각 좌표 방향의 변화율을 따로 계산한다

한 좌표만 바꾼다

다른 좌표는 고정하고 한 좌표에만 h 를 더한다

변화 방향	바꾼 입력	고정한 값
θ_1 방향	$(\theta_1 + h, \theta_2)$	θ_2
θ_2 방향	$(\theta_1, \theta_2 + h)$	θ_1

편미분(partial derivative)은 다른 좌표를 고정하고 한 좌표만 바꿀 때 함수값이 변하는 비율이다

Source: Zhang et al. (2023), Partial Derivatives and Gradients.

좌표별 편미분

$J(\theta_1, \theta_2) = (\theta_1 - 2)^2 + \theta_2^2$ 를 좌표별로 미분한다

$$\frac{\partial J}{\partial \theta_1} = 2(\theta_1 - 2)$$
$$\frac{\partial J}{\partial \theta_2} = 2\theta_2$$

미분하지 않는 좌표는 상수로 둔다

Source: Zhang et al. (2023), Partial Derivatives and Gradients.

그래디언트는 편미분을 모은 벡터

각 매개변수에 대한 편미분을 매개변수와 같은 순서로 쌓은 벡터가 그래디언트(gradient)

$$\nabla_{\theta} J(\theta) = \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} 2(\theta_1 - 2) \\ 2\theta_2 \end{bmatrix}$$

그래디언트의 각 성분은 해당 매개변수에 대한 함수값의 국소 변화율이다

그래디언트와 매개변수의 모양

입력	출력	미분 결과	모양
수 θ	수 J	$dJ/d\theta = J'(\theta)$	수 하나
벡터 θ	수 J	$\nabla_{\theta} J$	θ 와 같다
행렬 Θ	수 J	$\nabla_{\Theta} J$	Θ 와 같다

수 하나를 출력하는 함수의 그래디언트는 매개변수와 같은 모양이다

두 변수 이차함수

$\theta = (4, 1)$ 에서 그래디언트를 계산한다

$$J(4, 1) = (4 - 2)^2 + 1^2 = 5$$
$$\nabla J(4, 1) = \begin{bmatrix} 2(4 - 2) \\ 2(1) \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

θ_1 방향의 변화율 절댓값이 θ_2 방향보다 크다

중앙차분으로 그래디언트를 확인한다

중앙차분(central difference)은 한 좌표의 양쪽 함수값 차를 2ε 로 나누어 편미분을 근사한다

$$d_j = \frac{J(\theta + \varepsilon \mathbf{e}_j) - J(\theta - \varepsilon \mathbf{e}_j)}{2\varepsilon}, \quad \mathbf{e}_j = j\text{번째 성분만 1인 벡터}$$

바꾼 좌표	앞쪽 함수값	뒤쪽 함수값	d_j
θ_1	$J(4.1, 1) = 5.41$	$J(3.9, 1) = 4.61$	4
θ_2	$J(4, 1.1) = 5.21$	$J(4, 0.9) = 4.81$	2

J 가 이차함수라 $\varepsilon = 0.1$ 에서도 $(d_1, d_2) = (4, 2)$ 로 오차가 없다

대칭이 아닌 행렬의 이차함수

행렬의 두 비대각 원소가 한 교차항에 함께 들어간다

$$A = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}, \quad \theta = \begin{bmatrix} u \\ v \end{bmatrix}, \quad \mathbf{b} = \mathbf{0}$$

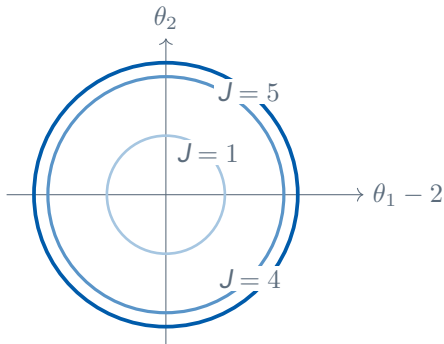
$$q(\theta) = \frac{1}{2} \theta^\top A \theta + \mathbf{b}^\top \theta = \frac{1}{2} (2u^2 + (1+3)uv + 4v^2) = u^2 + 2uv + 2v^2$$

$$\nabla q = \begin{bmatrix} 2u + 2v \\ 2u + 4v \end{bmatrix} = \frac{A + A^\top}{2} \theta + \mathbf{b}, \quad \nabla q(1, 2) = \begin{bmatrix} 6 \\ 10 \end{bmatrix}$$

A가 대칭이 아니면 $A\theta$ 만으로는 그래디언트를 구할 수 없다

같은 함수값을 잇는 등고선

J 값이 같은 점 (θ_1, θ_2) 를 이은 선이 등고선(contour)이다

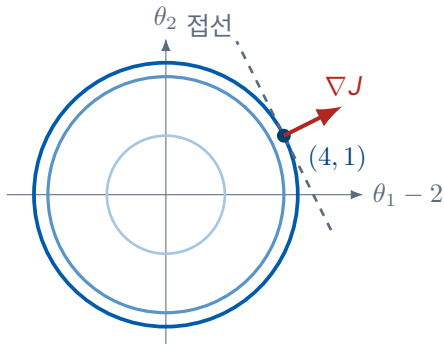


안쪽부터 $J = 1, 4, 5$ 인 등고선이며 중심 $\theta = (2, 0)$ 에서 멀수록 J 가 크다

Adapted from Zhang et al. (2023), Geometry of Gradients and Gradient Descent.

그래디언트와 등고선

$\theta = (4, 1)$ 에서 그래디언트 방향을 등고선 위에 표시한다



그래디언트는 등고선의 접선과 직교하며 함수값이 가장 빠르게 증가하는 방향이다

Adapted from Zhang et al. (2023), Geometry of Gradients and Gradient Descent.

작은 변화의 근사

$\Delta\theta$ 가 작을 때 내적으로 함수값의 변화를 근사한다

$$\Delta J \approx \nabla J(4, 1)^\top \Delta\theta$$

$$\nabla J(4, 1) = [4, 2]^\top$$

$$\Delta\theta = [-0.2, -0.1]^\top$$

$$\Delta J \approx 4(-0.2) + 2(-0.1) = -1.0$$

$$\text{실제 } J(3.8, 0.9) - J(4, 1) = 4.05 - 5 = -0.95$$

그래디언트와 반대 방향으로 조금 이동하면 1차 근사에서 함수값이 감소한다

Source: MIT 6.390 (2026), Matrix Calculus.

Checkpoint

Multivariable Gradients

- $\theta = (4, -1)$ 에서 그래디언트는 무엇인가
- 두 좌표 가운데 변화율의 절댓값이 큰 쪽은 어디인가
- J 를 감소시키는 작은 변화 $\Delta\theta$ 의 예는 무엇인가

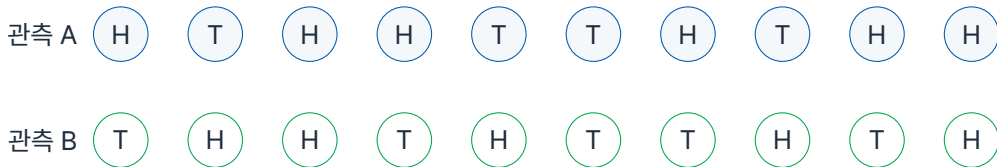
Section 3 • Derivatives and Probability

Uncertainty in Observations

표본공간, 사건, 확률변수, 통계량의 구분

같은 시행, 다른 결과

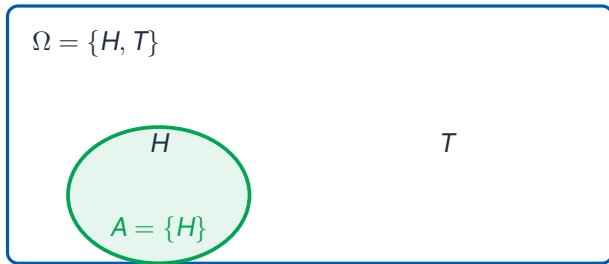
같은 조건에서 동전을 열 번씩 던져도 결과는 매번 달라질 수 있다



관측 A는 앞면이 6번, 관측 B는 5번 나와 순서와 개수가 모두 다르다

표본공간과 사건

표본공간(sample space)은 가능한 결과 전체이고, 사건(event)은 그 부분집합이다



$\Omega = \{H, T\}$ 에서 사건 $A = \{H\}$ 는 앞면이 나오는 경우이다

확률분포의 기본 규칙

확률분포(probability distribution)는 가능한 결과에 확률을 배정한다

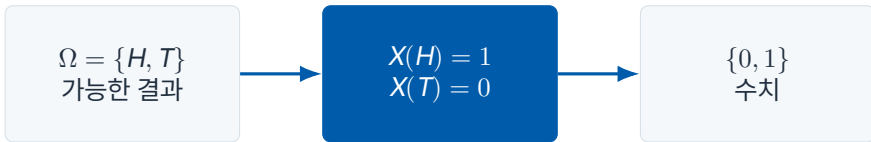
조건	식	동전 예
0 이상	$P(\omega) \geq 0$	$P(H), P(T) \geq 0$
전체 확률	$\sum_{\omega \in \Omega} P(\omega) = 1$	$P(H) + P(T) = 1$
사건의 확률	$P(A) = \sum_{\omega \in A} P(\omega)$	$P(\{H\}) = P(H)$

가능한 모든 결과에 배정한 확률의 합은 1이다

Source: Zhang et al. (2023), A More Formal Treatment.

확률변수는 결과를 수로 바꾼다

확률변수(random variable)는 각 결과에 수 하나를 대응시킨다



확률변수 X 는 앞면을 1, 뒷면을 0으로 바꾸는 함수이다

베르누이 확률변수

베르누이(Bernoulli) 확률변수는 두 결과를 0과 1로 기록한다

결과	X	확률	매개변수
H	1	p	$P(X = 1) = p$
T	0	$1 - p$	$P(X = 0) = 1 - p$

베르누이 분포의 매개변수 p 는 $X = 1$ 이 될 확률이다

동전을 열 번 던진 결과

앞면을 1, 뒷면을 0으로 기록한 열 번의 관측값을 사용한다

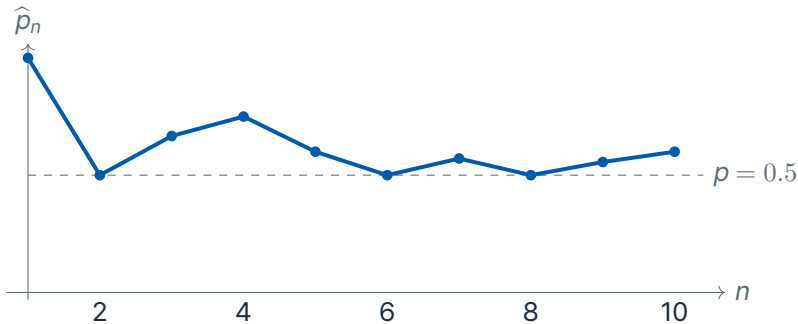


$$\sum_{i=1}^{10} X_i = 6 \quad \hat{p} = \bar{X} = \frac{6}{10} = 0.6$$

표본평균(sample mean)은 앞면의 상대도수 $6/10 = 0.6$ 과 같다

상대도수로 확률을 추정한다

시행 수 n 에 따라 누적 상대도수 $\hat{p}_n$ 이 어떻게 변하는지 확인한다



시행 횟수가 적을 때는 한 번의 결과가 추정값에 미치는 영향이 크다

Adapted from Zhang et al. (2023), A Simple Example: Tossing Coins.

매개변수와 통계량

대상	기호	뜻
매개변수	p	시행 과정의 확률, 직접 관측되지 않음
관측값	X_i	한 번의 시행에서 기록한 0 또는 1
통계량	$\hat{p}_n = \bar{X}_n$	n 개 관측값으로 계산한 추정값

매개변수는 확률분포의 속성이고, 통계량은 관측 자료에서 계산한 값이다

Source: Zhang et al. (2023), A Simple Example: Tossing Coins.

Checkpoint

Uncertainty in Observations

- 동전 던지기의 표본공간과 사건 $\{H\}$ 는 무엇인가
- 확률변수 X 는 각 결과에 어떤 수를 대응시키는가
- 매개변수 p 와 통계량 $\hat{p}_{10}$ 은 어떻게 다른가

Expectation and Variance

기댓값과 분산으로 확률분포를 요약

기댓값은 확률 가중평균

기댓값(expectation)은 각 값과 그 확률의 곱을 모두 더한 분포의 중심이다

$$\mathbb{E}[X] = \sum_x xP(X = x)$$

x	확률변수 X 가 가질 수 있는 값
$P(X = x)$	값 x 가 나타날 확률
$xP(X = x)$	값 x 가 기댓값에 기여하는 양

Source: Zhang et al. (2023), Probability and Statistics: Expectations.

기댓값 계산 예

세 값에 각각의 확률을 곱한 뒤 모두 더한다

x	$P(X = x)$	$xP(X = x)$
-1	0.25	-0.25
0	0.50	0
3	0.25	0.75
합계		0.50

베르누이 확률변수의 기댓값은 $\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$ 이다

기댓값 0.5는 분포의 중심이며 실제로 가능한 값일 필요는 없다

분산은 제곱편차의 기댓값

분산(variance)은 각 값이 기댓값에서 떨어진 정도를 제공한 값의 기댓값이다

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

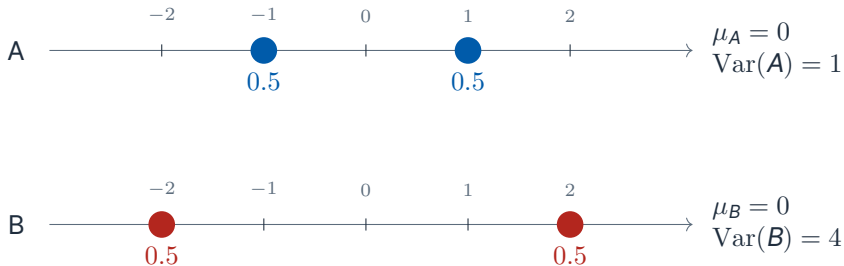
편차	$X - \mathbb{E}[X]$
제곱편차	부호를 없애고 큰 차이를 더 크게 반영
분산	제곱편차의 기댓값

앞의 분포에서 $\mathbb{E}[X^2] = 2.5$ 이므로 $\text{Var}(X) = 2.5 - 0.5^2 = 2.25$

Source: Zhang et al. (2023), Probability and Statistics: Expectations.

기댓값이 같아도 분산은 다를 수 있다

두 분포의 기댓값은 모두 0이지만 값이 퍼진 정도는 다르다



기댓값만으로 분포의 퍼진 정도를 구분할 수 없다

표준편차

표준편차(standard deviation)는 분산의 제곱근으로 원자료와 단위가 같다

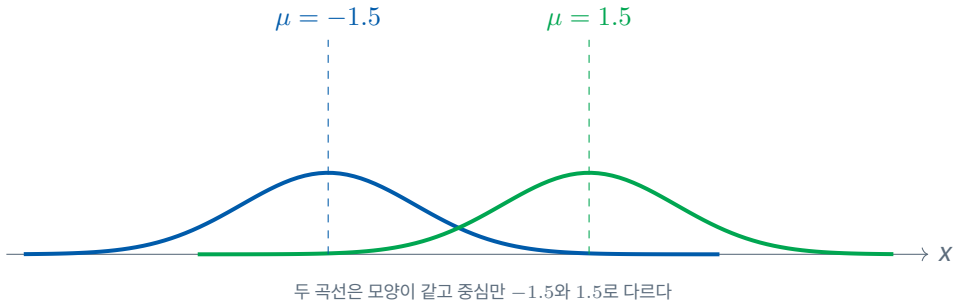
요약값	정의	단위
기댓값(평균 μ)	$\mu = \mathbb{E}[X]$	X 와 같음
분산	$\sigma^2 = \text{Var}(X)$	X 단위의 제곱
표준편차	$\sigma = \sqrt{\text{Var}(X)}$	X 와 같음

표준편차는 값이 퍼진 정도를 원자료와 같은 단위로 나타낸다

Source: Zhang et al. (2023), Random Variables: Standard Deviations.

가우스 분포의 평균

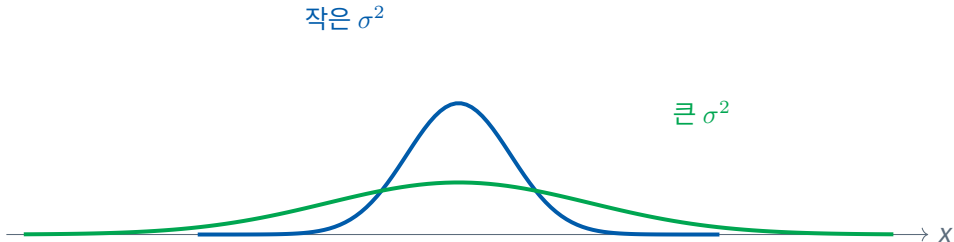
가우스 분포(Gaussian distribution)에서 분산을 고정하고 평균만 바꾸면 곡선의 중심이 이동한다



Adapted from Zhang et al. (2023), Distributions: Gaussian.

가우스 분포의 분산

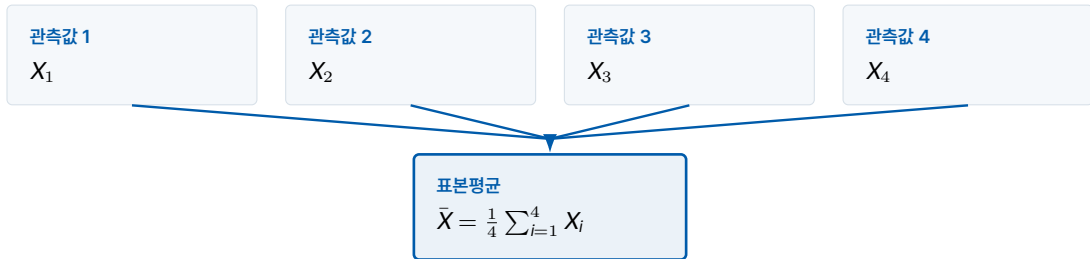
평균을 고정하고 분산을 바꾸면 곡선이 퍼진 정도가 달라진다



분산이 클수록 확률밀도 곡선이 넓게 퍼지고, 곡선 아래 넓이가 확률이므로 두 곡선의 전체 넓이는 모두 1이다

표본평균에서 한 관측값의 비중

관측값 네 개를 평균하면 각 관측값의 계수는 $1/4$ 이다



관측값 하나가 δ 만큼 변하면 표본평균은 $\delta/4$ 만큼 변한다

표본평균의 분산

$X_1, \dots, X_n$ 이 서로 독립이고 분산이 모두 σ^2 인 경우이다

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$
$$\text{Var}(\bar{X}) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}$$

$1/n^2$ 은 $\text{Var}(cX) = c^2 \text{Var}(X)$ 에서, 합의 분산이 분산의 합이 되는 것은 서로 독립이라는 조건에서 나온다

서로 독립인 관측값 n 개의 표본평균의 분산은 개별 관측값 분산의 $1/n$ 배이다

Source: Zhang et al. (2023), Random Variables: Variances.

함수의 국소 변화와 자료의 변동

개념	설명하는 대상	요약
도함수	한 매개변수의 변화	함수값의 국소 변화율
그래디언트	매개변수 벡터의 변화	좌표별 국소 변화율 벡터
기댓값	확률분포의 중심	확률 가중평균
분산	중심 주변의 변동	제곱편차의 기댓값

도함수와 그래디언트는 함수의 국소 변화를, 기댓값과 분산은 분포의 중심과 퍼짐을 나타낸다

Checkpoint

Expectation and Variance

- x 와 $P(X = x)$ 로 기댓값을 계산하는 식은 무엇인가
- 분산과 표준편차의 단위는 어떻게 다른가
- 서로 독립인 관측값의 표본평균의 분산이 σ^2/n 인 이유는 무엇인가

Why Does a Model Fit Training Data but Fail on New Data?

Next Lecture: Learning and Generalization

- 학습 자료와 새 자료에서의 성능
- 과적합의 의미
- 편향과 분산