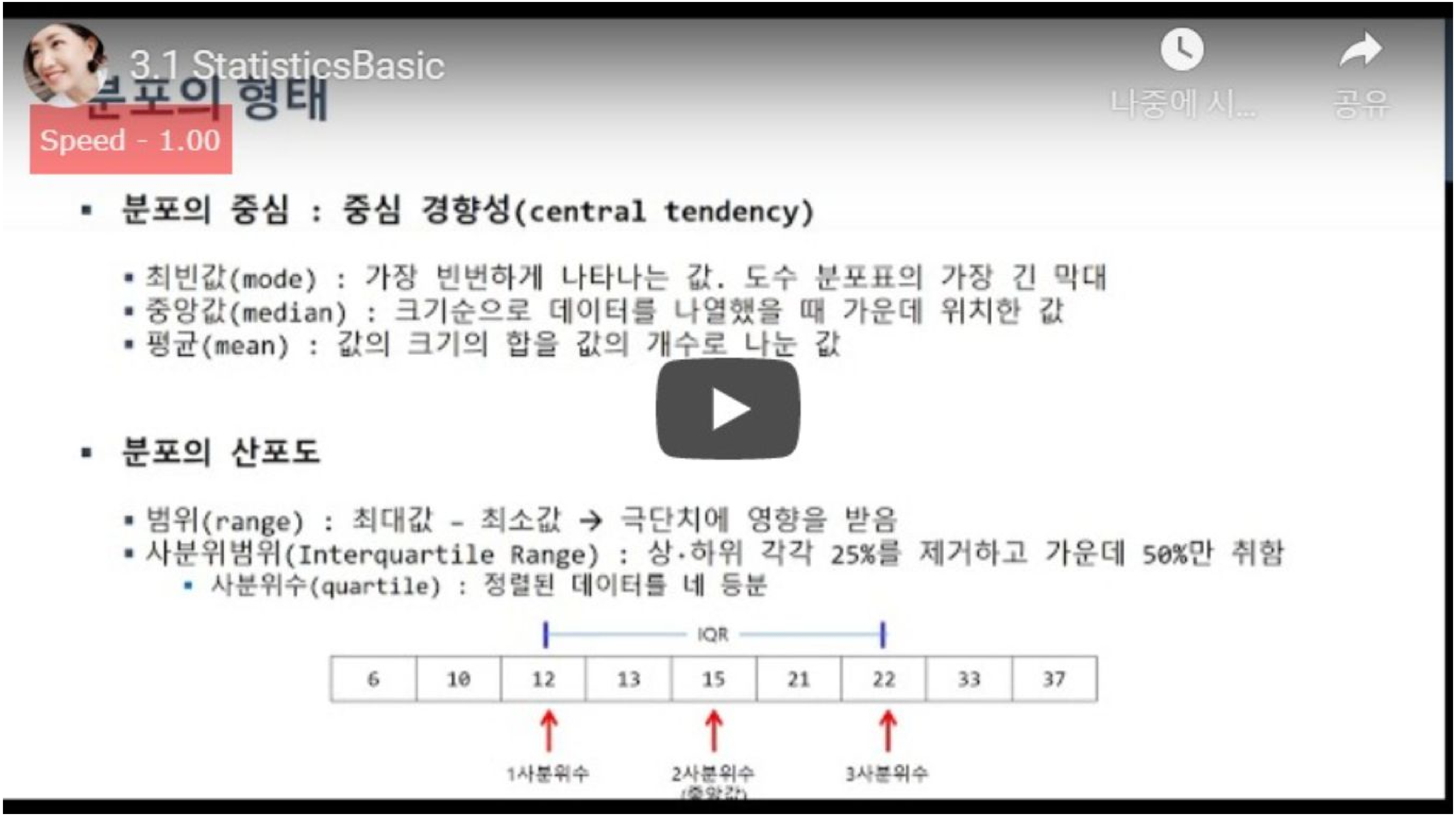
제목없음


3.1 StatisticsBasic
분포의 형태
Speed - 1.00
나중에 시...
공유
▪ 분포의 중심 : 중심 경향성(central tendency)
▪ 최빈값(mode) : 가장 빈번하게 나타나는 값. 도수 분포표의 가장 긴 막대
▪ 중앙값(median) : 크기순으로 데이터를 나열했을 때 가운데 위치한 값
▪ 평균(mean) : 값의 크기의 합을 값의 개수로 나눈 값
▪ 분포의 산포도
▪ 범위(range) : 최대값 - 최소값 → 극단치에 영향을 받음
▪ 사분위범위(Interquartile Range) : 상·하위 각각 25%를 제거하고 가운데 50%만 취함
▪ 사분위수(quartile) : 정렬된 데이터를 네 등분
IQR
6 10 12 13 15 21 22 33 37
1사분위수
2사분위수
3사분위수

예제로 알아보는 모델링과 예측

Speed - 3.10

- **영업 사원의 월급**
  - 자동차 판매회사의 신입 사원이 다음과 같이 계약

    100만원 기본급에 자동차를 1대 팔 때마다 90만원을 추가로 받는다.

  - 이 계약 조건을 기반으로 모델링
    - 판매 대수를 x, 월급을 y라 하면
    - 수식으로 표현하면

    $$y = 1000000 + 900000x$$

  - 위 수식을 **모델**이라 부름
  - 변수를 뽑고 변수 사이의 관계를 나타내는 수식을 구하는 과정을 **모델링**이라 부름

  - 모델이 있으면 예측이 가능
    - 다음 달에 3대를 팔면 월급이 얼마일까? → 370만원
    - 더욱 분발하여 그 다음 달에 20대를 팔면? → 1900만원

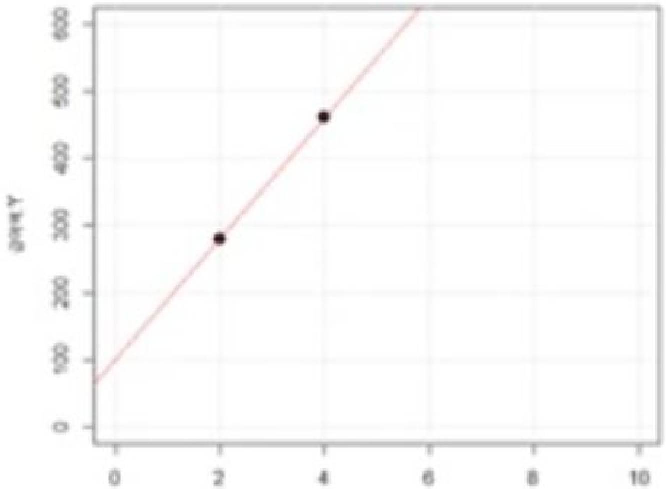
예제로 알아보는 모델링과 예측

- 조건이 아닌, 데이터로부터 모델을 생성
  모델 구축에 사용하는 데이터를 훈련 집합(training set)이라 함
- $X = \{x_1, x_2, \dots, x_n\}$, $Y = \{y_1, y_2, \dots, y_n\}$
  - (x_i, y_i) 는 i번째 관측(observation) 또는 i번째 샘플(sample)
  - x_i는 특징(feature)
  - y_i는 레이블(label), ground truth(GT). 즉, 정답
  - X는 독립변수·설명변수(explanatory variable)
    Y는 종속변수·반응변수(response variable)
- x_i가 입력되면 y_i를 맞추는 문제
- 모델링이란, 훈련 집합을 이용하여 최적의 모델을 찾아내는 과정

예제로 알아보는 모델링과 예측

- 영업 사원의 예에서, 한 사원이 급여 조건을 몰랐다고 가정
- 첫 달에 2대를 팔고 280만원, 둘째 달에 4대를 팔아 460만원을 받음
- 샘플 데이터는 $X_{판매대수}$ = {2, 4}, $Y_{수령액}$ = {2800000, 4600000}
- $Y = \alpha_1 X + \alpha_0$ 식에 훈련 데이터를 대입하여, α_0, α_1을 계산하면

$$Y = 900000X + 1000000$$ ← 모델



- 훈련 집합을 시각화하고,
  X, Y 변수로 구축한 모델은 빨간 선
- α_0, α_1 을 매개변수(parameter)라 하고,
  최적의 매개변수 값을 알아내는 과정을 모델 적합

예제로 알아보는 모델링과 예측

Speed - 3.10

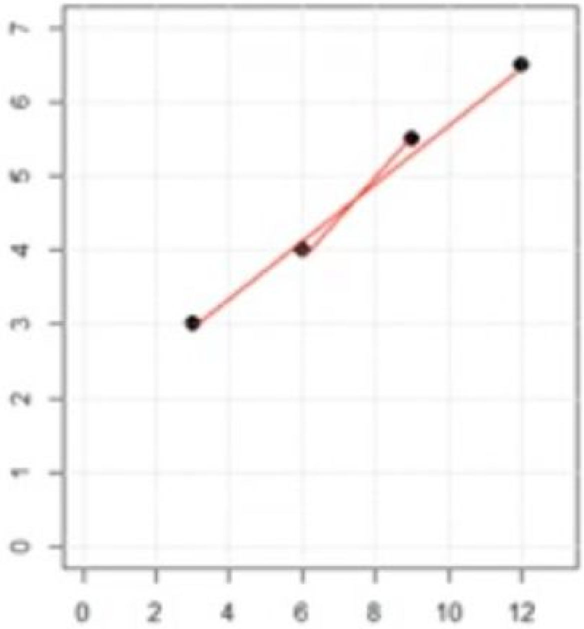
- 방정식의 수립 : 모델 선택(model selection)
  - $Y = \alpha_1 X + \alpha_0$
  - α_0, α_1 → 매개변수(parameter)
- 최적의 매개변수 값을 알아내는 과정 : 모델 적합(model fitting), 학습(learning), 훈련(training)
- 모델을 이용하여 훈련 집합에 없는 새로운 샘플에 대하여 예측
  - x=5 (판매 대수 5대) 를 모델 $Y = 900000X + 1000000$ 에 대입하여 y = 5500000
  - 즉, 550만원의 급여액 추정

모델의 품질 평가

- **실제 측정 데이터는 불확실성과 측정 오차가 존재**
  - 체온의 경우, 수 분 단위로 변동하고 옥수수 한 개의 낱알 수는 나무와 줄기마다 차이

- **실험을 통해 측정한 데이터**
  - 전기량 x에 따른 물체의 이동거리 y를 측정

X	Y
3.0	3.0
6.0	4.0
9.0	5.5
12.0	6.5



모델의 품질 평가

Speed - 3.40

- 모델 선택 → 선형 방정식
- 모든 샘플이 한 직선상에 존재하지 않음

- 방법 1. 선형 모델 대신 2차, 3차의 고차 방정식 채택
  → 복잡한 모델은 과잉적합(overfitting)의 위험
- 방법 2. 선형 모델을 사용하되, 오차를 허용
  → 오차 0은 현실적으로 불가능. 최소한의 오차를 허용한 최적의 모델을 찾음

모델의 품질 평가

- 모델(1) $y = 0.5x + 1.0$

x_i	3.0	6.0	9.0	12.0
예측값 $f(x_i)$	2.5	4.0	5.5	7.0
관측값 GT y_i	3.0	4.0	5.5	6.5
오차	0.5	0.0	0.0	-0.5

- 평균 제곱 오차(MSE)Mean Squared Error

  $\frac{1}{n}\sum_1^n\left(y_i - f(x_i)\right)^2$

  = 0.125

- 모델(2) $y = 5/12x + 7/4$

x_i	3.0	6.0	9.0	12.0
예측값 $f(x_i)$	3.0	4.25	5.5	6.75
관측값 GT y_i	3.0	4.0	5.5	6.5
오차	0.0	-0.25	0.0	-0.25

- 평균 제곱 오차(MSE)Mean Squared Error

  $\frac{1}{n}\sum_1^n\left(y_i - f(x_i)\right)^2$

  = 0.03125

- 모델(2)가 MSE가 더 낮으므로 채택
- MSE를 최소화하는 모델을 찾는 것 → 모델의 최적화 문제(Optimization problem)

R을 이용한 모델 적합

- 훈련 데이터

X	Y
3.0	3.0
6.0	4.0
9.0	5.5
12.0	6.5

```
> x <- c(3.0, 6.0, 9.0, 12.0)
> y <- c(3.0, 4.0, 5.5, 6.5)
> m <- lm(y ~ x)
> m

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
       1.75         0.40
```

α_0 α_1

- 모델 : $Y = 0.4X + 1.75$