

학습계획서

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제	주요내용
1일차 (5 / 27)	황성욱	- 오리엔테이션	향후 학습방향 조정 및 교육컨텐츠 선정
2일차 (5 / 28)	최재림	- 자료의 형태와 요약1, 2	1-1. 자료(변수)의 두 가지 형태 1-2. 범주형 자료의 요약 1-3. 양적 자료의 요약 2-1. 대표값 2-2. 산포도 2-3. 사분위범위 2-4. 상자그림
3일차 (5 / 29)	황성욱	- 확률변수와 분포	1. 확률과 임의성 2. 이산확률변수 3. 연속확률변수 4. 평균과 분산
4일차 (5 / 30)	최재림	- 정규분포	1. 정규분포의 개념 2. 정규분포의 형태 3. 정규분포의 표준화
5일차 (5 / 31)	황성욱	- 표본분포와 중심극한정리	1. 모집단과 표본 2. 표본분포 3. 중심극한정리
6일차 (6 / 3)	최재림	- 통계적 추론 및 검정	1-1. 통계적 추론 1-2. 통계적 추정 1-3. 표본크기의 결정 1-4. t-분포 1-5. 일표본 t-신뢰구간 2-1. 통계적 검정의 개념 2-2. 두 종류의 가설 2-3. 두 종류의 오류 2-4. P값 2-5. P값을 이용한 유의성 검정의 단계 2-6. 단측검정과 양측검정
7일차 (6 / 4)	황성욱	- 모평균에 대한 검정	1. 모평균의 검정: Z-검정 2. 검정통계량 3. 양측검정과 단측검정의 경우 P값 4. 신뢰구간과 가설검정 5. 일표본 t-검정 6. 유의성 검정에 대한 주의점 7. 정규분포가 아닐 때의 추론
8일차 (6 / 5)	최재림	- 상관분석	1. 상관분석의 개념 2. 상관계수 3. 상관계수 행렬
9일차 (6 / 7)	황성욱	- 단순선형 회귀분석	1. 단순선형회귀분석의 개념 2. 회귀직선의 적합 3. 최소제곱회귀직선 4. Y값의 예측 : 내삽법(보간법)과 외삽법(보외법) 5. 결정계수 6. 변수변환 7. 회귀계수의 검정 8. 회귀직선의 유의성검증 9. 잔차의 분석
10일차 (6 / 10)	최재림	- 분산분석	1. 분산분석의 개념 2. 일원분류분산분석의 모형 3. ANOVA F-검정 4. 분산분석표

학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제
1일차 (5 / 27)	황성욱	- 오리엔테이션

주요 내용 요약

1. 학습 주제 선정
2. 팀 구성
3. 빅데이터 학습 플랫폼 선정
4. 구체적인 학습 계획 수립

학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제
2일차 (5 / 28)	최재림	- 자료의 형태와 요약1, 2

주요 내용 요약

자료의 형태와 요약 I

자료(변수)의 두 가지 형태

- 1) categorical(범주형): 명목/순서
- 2) quantitative(양적): 연속/이산

- 명목(Nominal) 변수: 순서 없는 범주를 가지는 변수
예) 성별(남, 여), 지역(서울, 부산, 광주...)
- 순서(Ordinal) 변수: 순서가 있는 범주를 가지는 변수
예) 자동차 크기(소형, 중형, 대형), 계층(상, 중, 하)
- 연속(Continuous) 변수: 무수히 많은 다른 값을 가짐
예) 키, 몸무게, 온도
- 이산(Discrete) 변수: 몇 개의 다른 값만 가짐
예) 고장 횟수, 가족 구성원의 수

자료는 범주형/양적 자료로 나누어 지며, 각 자료형을 요약하기 위해서 사용되는 기법은 다음과 같다.

1. 범주형

- 도수분포표
- 막대그래프
- 파이 차트

2. 양적

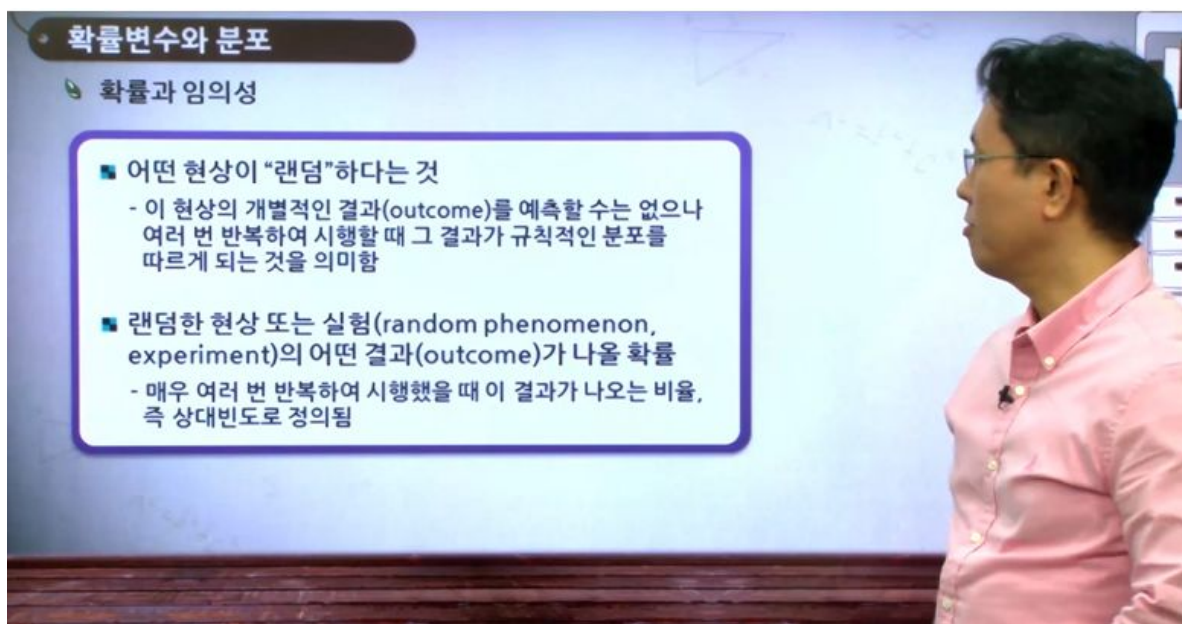
- Graphical 요약
Dotplot, Stemplot, Histogram, Boxplot, Linegraph 등등..
- 수치적 요약
대표값(산술평균, 중앙값, 최빈값), 산포도(범위, 사분위범위, 표준편차)

학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제
3일차 (5 / 29)	황성욱	- 확률변수와 분포

주요 내용 요약



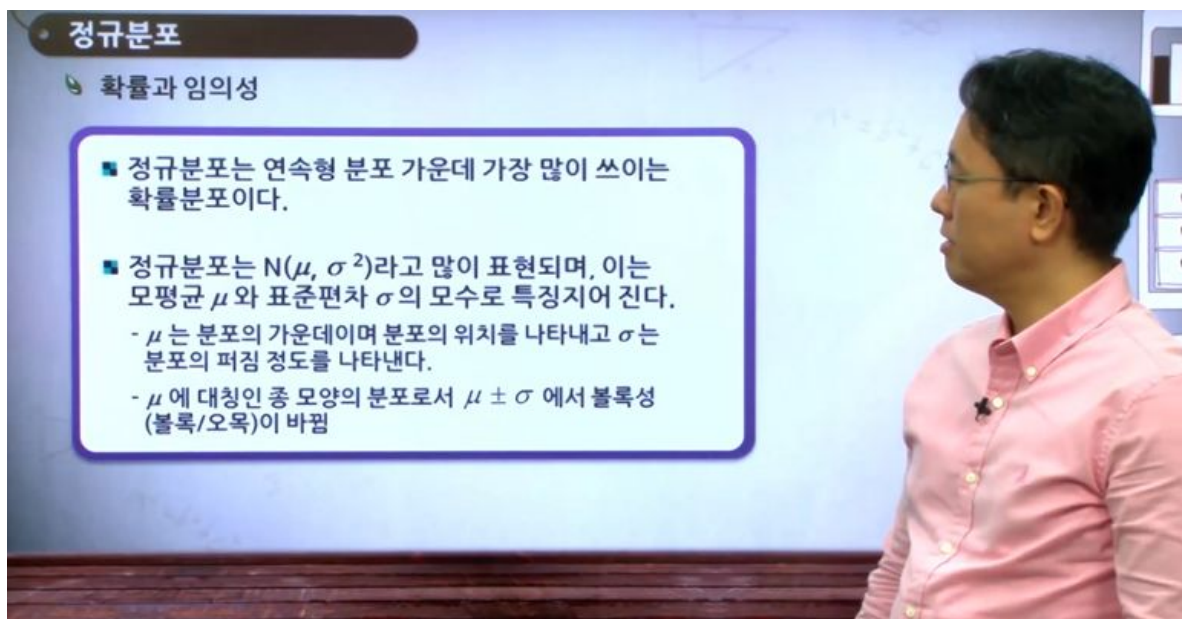
- 확률과 임의성 = 현상이 랜덤하다
- 확률변수 = 현상 또는 실험의 결과로 결정되는 수치적인 양 (numerical quantity)
 일정한 확률분포를 가짐(이산형/연속형)
- ex) 동전던지기
 동전던지기의 결과는 랜덤하지만 각 시행(던지기)이 독립적이라는 가정하에 여러 번 던졌을 때의 결과는 예측 가능하다
 X = 각 시행에서 앞면이 나오는 횟수 (확률변수)
 $P(X=1) = p$
 $P(X=0) = 1-p$ (확률분포)
- 이산확률변수 = 유한 또는 셀 수 있는 무한의 값만을 가질 수 있음
- 연속확률변수 = 어떤 구간 안의 모든 값을 다 취할 수 있는 변수
- 평균과 분산

학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제
4일차 (5 / 30)	최재림	- 정규분포

주요 내용 요약



-정규분포의 개념 =

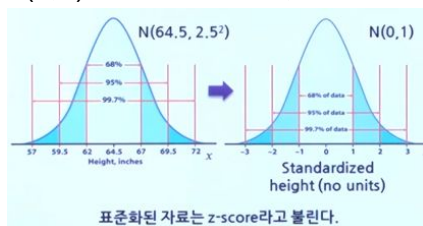
정규분포는 연속형 분포 가운데 가장 많이 쓰이는 확률 분포이다.
모평균과 표준편차의 모수로 특징지어 진다.

-표준 정규분포 = 모평균:0, 표준편차:1인 정규분포

-정규분포의 형태 = 표준편차가 작은 경우에는 평균 주위에 가깝게 몰려있게되고,
표준편차가 큰 모집단의 분포는 넓게 퍼져있는 형태를 취한다.

-정규분포의 표준화 =

모든 정규분포는 같은 형태적 성질을 갖기 때문에 $N(\mu, \sigma^2)$ 를 표준화해 표준 정규분포 $N(0,1)$ 을 얻을 수 있고, 표준화 후 $N(0,1)$ 의 확률표를 이용해 확률계산을 할 수 있다.



학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제
5일차 (5 / 31)	황성욱	- 표본분포와 중심극한정리

주요 내용 요약

표본분포와 중심극한정리

모집단과 표본

- 모집단 (population)**
 - 어떤 연구에서 실제 관심 있는 집단으로 흔히 전체를 모두 연구하기 어려움
 - 예: 모든 인간, 전국의 모든 근로자, 전국의 모든 유권자, 모든 금붕어, ...
- 표본 (sample)**
 - 모집단의 일부분으로서 실제로 연구자가 자료를 수집하여 연구하는 부분
 - 표본추출이 잘 되어야 연구전체가 의미 있어짐
- 모수 (parameter)**
 - 모집단의 특성을 나타내는 숫자
 - 미지의 고정된 상수
- 통계량 (statistic)**
 - 표본의 특성을 나타내는 숫자
 - 표본에 따라 다른 값을 갖는 확률변수
 - 모수를 추정하는 데에 사용됨

모집단 (모집단 안에 표본이 포함됨)

1. 모집단과 표본

모집단 : 정보를 얻고자 하는 관심 대상의 전체 집합

표본 : 모집단의 부분집합

2. 표본분포 = 표본에서 도출되는 통계량에 대한 확률분포

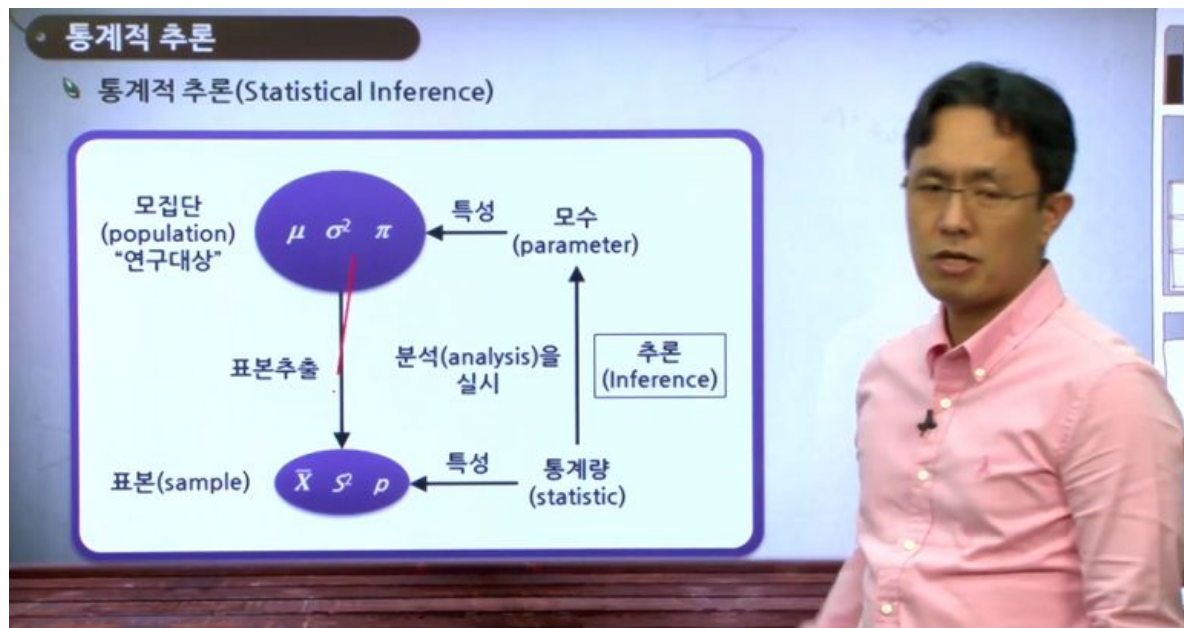
3. 중심극한정리 = 표본의 갯수(N)이 충분하다면 모수를 모르는 상황에서도 표본 통계량으로 정규분포를 구성하여 모수를 추정할 수 있다는 것

학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제
6일차 (6 / 03)	최재림	- 통계적 추론 및 검정

주요 내용 요약



1-1. 통계적 추론

모집단에 대한 어떤 미지의 양상을 알기 위해 통계학을 이용하여 추측하는 과정

1-2. 통계적 추정

표본조사를 토대로 모집단의 성질을 추정하는 작업

1-3. 표본크기의 결정

신뢰수준 : Z

모집단 분사의 추정치 : σ^2

허용오차 : d

$$n = Z^2 \cdot \sigma^2 / d^2$$

1-4. t-분포

1-5. 일표본 t-신뢰구간

2-1. 통계적 검정의 개념

2-2. 두 종류의 가설

2-3. 두 종류의 오류

2-4. P값

2-5. P값을 이용한 유의성 검정의 단계

2-6. 단측검정과 양측검정

단측검정과 양측검정은 기본적으로 같은 검정인데, 단지 보수적인(conservative) 정도(degree)에 차이가 있다.

물론, 주장하고자 하는 대립가설에 따라 단측검정을 하거나 양측검정을 하므로 차이가 있지만, 같은 검정이라고 생각해도 큰 무리는 없다. (즉, 귀무가설은 '=')

학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제
7일차 (6 / 04)	황성욱	- 모평균에 대한 검정

주요 내용 요약

모평균에 대한 검정

모평균의 검정: Z-검정

- $H_0: \mu = \mu_0$ 의 가설을 모평균이 μ (unknown)이고 표준편차가 σ (known)인 정규분포에서 뽑힌 크기 n 인 랜덤표본으로부터 검정하고자 할 때, 표본평균의 분포가 $N(\mu, \sigma^2/n)$ 를 따름을 이용한다.
- P-value는 귀무가설이 맞다는 가정하에서 표본으로부터 얻은 관측치만큼 또는 그보다 더 극단적인(대립가설의 방향으로) 관측치가 얻어질 확률이다.
- 검정통계량:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- 모평균의 검정: Z-검정 = 제 1종의 오류의 최대허용확률로 유의수준이 결정되면 기각역이 결정됨
- 검정통계량 = 검정통계량은 귀무가설이 참일 때 기대되는 값과 관측치 사이의 차이를 측정 (관측치가 가정된 값으로부터 몇 표준편차만큼 떨어져 있는가?)
- 양측검정과 단측검정의 경우 P값
- 신뢰구간과 가설검정
신뢰구간은 귀무가설의 기각 또는 채택이라는 black and white의 답만을 제공하지만, 모평균 μ 가 가질 가능성이 있는 값들을 추정하기도 한다
- 일표본 t-검정 = 하나로 구성된 모집단의 평균 값을 기준 값과 비교하고자 할 때 사용되는 분석법
- 유의성 검정에 대한 주의점
유의 수준 α 를 결정할 때
고려할 점 : 귀무가설을 기각했을 때 어떤 조치가 취해질 것인가
Preliminary study를 하고 있다면, 어떤 의미 있는 결과를 놓치지 않도록 조금 큰 α 의 값을 사용할 필요가 있다
관례적으로 : 업계 또는 학계의 standard를 주로 사용하게 됨
cutoff point를 생각하기 어렵다
p-value 크기에 따라 “약간 유의”, “유의” 또는 “매우 유의”하다고 표현할 수 있다
- 정규분포가 아닐 때의 추론 = 모집단의 분포가 정규분포와 다르고 표본의 크기가 작으면 정규성 가정을 할 수 없다

학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제
8일차 (6 / 05)	최재림	- 상관분석

주요 내용 요약

상관분석
상관분석(Correlation analysis)

- 양의 상관(positive correlation): 한 변수(X)의 값이 증가하면 다른 변수(Y)의 값도 증가한다.
- 예: 나이가 증가할수록 혈압이 증가한다.
- 음의 상관(negative correlation): 한 변수(X)의 값이 증가하면 다른 변수(Y)의 값은 감소한다.
- 예: 나이가 증가할수록 기억력이 감소한다.
- 상관이 영(zero): 두 변수 사이에 선형적인 관련이 없다.
- 상관계수는 Pearson's correlation coefficient로 추정한다.

1. 상관분석의 개념

두 변수간에 어떤 선형적 관계를 갖고 있는 지를 분석하는 방법

2. 상관계수

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- 특징

- 1) 상관계수는 단위와 관계가 없다.
- 2) 상관계수는 -1부터 1까지의 값을 갖는다.
- 3) 두 변수를 서로 바꿔도 상관계수의 값은 똑같다.

3. 상관계수 행렬

공분산 행렬, 상관계수 행렬 비교

공분산 행렬 =	X1	X2	X3	
	15.38	2.49	7.17	X1
	2.49	7.31	0.44	X2
	7.17	0.44	6.63	X3
상관계수 행렬 = $H_0: \rho=0$	X1	X2	X3	
	1.00	0.23	0.70	X1
	0.23	1.00	0.06	X2
	0.70	0.06	1.00	X3

학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------


일정	발제자	주제
9일차 (6 / 07)	황성욱	- 단순선형 회귀분석

주요 내용 요약

단순선형회귀분석

단순선형회귀분석

- 회귀분석이란 변수들간의 함수적인 관련성을 규명하기 위하여 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 자료로부터 추정하는 통계적 분석방법이다.
- 일반적으로 이 추정된 모형을 사용하여 필요한 예측을 하거나 관심 있는 통계적 추정과 검정을 실시한다.
- 단순 선형 회귀분석(simple linear regression)은 반응(종속)변수 y 와 하나의 설명(독립)변수 x 와의 선형적 관련성을 규명하는 회귀분석이다.
- 설명변수 (또는 독립변수)는 흔히 X 로 표현하며 반응변수에 영향을 주는 변수이고, 반응변수 (또는 종속변수)는 흔히 Y 로 표현하며 어떤 실험이나 조사의 결과를 측정하는 변수이다.



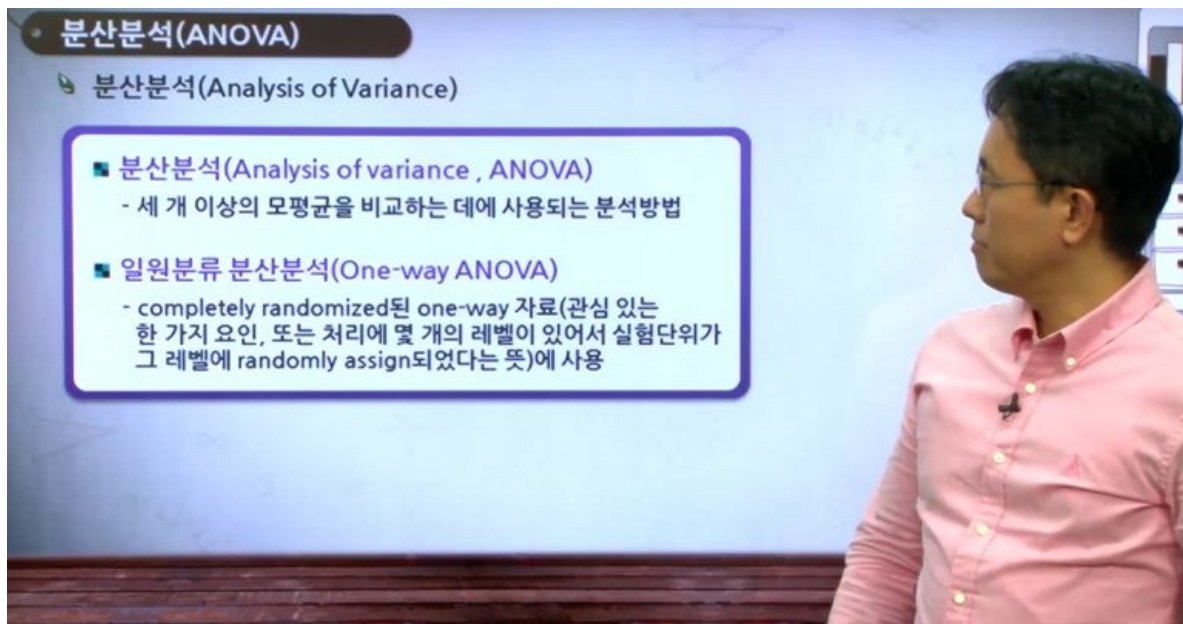
1. 단순선형회귀분석의 개념
회귀분석이란 변수들간의 함수적인 관련성을 규명하기 위하여 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 자료로부터 추정하는 통계적 분석방법
단순 선형 회귀분석(simple linear regression)은 반응(종속)변수 y 와 하나의 설명(독립)변수 x 와의 선형적 관련성을 규명하는 회귀분석
2. 회귀직선의 적합 : 자료를 가장 잘 설명해주는 직선을 찾는 것이 목적
3. 최소제곱회귀직선
4. Y 값의 예측 : 내삽법(보간법)과 외삽법(보외법)
5. 결정계수
 Y 의 변동 중에 x 에 대한 회귀식으로 설명되는 변동의 퍼센트
 X 와 Y 의 선형관계의 강약을 나타냄
단순선형회귀에서는 상관계수의 제곱과 정확히 일치
6. 변수변환
7. 회귀계수의 검정
8. 회귀직선의 유의성 검증
9. 잔차의 분석
회귀분석에서 제공되는 결과의 타당성 : 오차에 대한 기본가정의 검토가 필요

학습 정리

팀	슈퍼빅데이터	구성원	최재림, 황성욱
---	--------	-----	----------

일정	발제자	주제
10일차 (6 / 10)	최재림	- 분산 분석

주요 내용 요약



1. 분산분석의 개념

세 개 이상의 모평균을 비교하는 데에 사용되는 분석방법

2. 일원분류분산분석의 모형

관심 있는 한 가지 요인, 또는 처리에 몇 개의 레벨이 있어서 실험단위가 그 레벨에 randomly assign되어있는 자료에 사용

3. ANOVA F-검정

F-분포를 이용하여 집단 간의 분산과 집단 내의 분산을 비교한다.

4. 분산분석표