텍스트, 스크린샷, 편지, 폰트이(가) 표시된 사진

자동 생성된 설명

확률통계 및 회귀분석

챕터 1.1

통계학 – 무언가를 확률적으로 추론하는게 목표. Sample을 활용하여 Population 추론

(확률적 근거로 추론(설명)할거임!)

Population(모집단)

1. 우리가 샘플을 가지고 추론하려는 목표
2. Largest of random variables

Sample(표본)

1. A Part of population (Can estimate)

Statistical inference(확률적 추론) – population의 characteristic(특징)을 sample을 활용하여 추론하는 것. 따라서 질 좋은 샘플을 다량 확보하는 것이 매우 중요하다.

Sampling Error – 샘플의 수가 부족해서 생기는 오류

Non-sampling Error – 샘플을 따온 곳이 이상한 곳이라서 생기는 에러

챕터 1.2

Descriptive Statistics(기술 통계) – 걍 적어놓은 통계. 무언가를 추론하려는 목적이 아님.

Statistic parameters(유용한 통계 지표들)

1. Central tendency(중심 경향성) – mean, median, mode(최빈값)
2. Dispersion(산포도) –

range, variance, standard deviation(표준편차), Coefficient of Variance(CV, 변동계수), percentiles and quantiles(백분위, 분위수), interquartile range(사분위수)

디스퍼젼 Dispersion? 얼마나 퍼졌노?

Statistics(샘플에서 나온 측정값) –---(추론)---> Parameters(모집단의 기술통계값 ex)분산 표준편차 평균 등)

모집단을 추론하고 싶다. -> maybe Population is Normal distribution

1. 정규분포인건 알겠는데
2. 정규분포의 정확한 모양을 알고싶음!
3. 그러면 평균이랑 분산이 있어야겠네? (Parameters 요구)
4. 이 2개(파라미터)를 알면 정규분포를 따르는 모집단의 생김새를 추론할 수 있음

따라서 통계학이라는 것은 sample의 statistics를 활용하여 Population(모집단)의 Parameters를 추론하는 것.

**Parameters의 종류**

1. Location Parameter : 분포가 어디에 위치했는지를 보여줌(mean, median)
2. Dispersion Parameter : 분포의 산포도를 보여줌(Var, Spread)

샘플은 전체를 대변할 수 없음. 따라서 샘플을 활용하여 모집단을 추정하면 오차가 생길수밖에 없음. -> 신뢰구간 쓰면 되겠네?

신뢰구간 95%란, 내가 추정한 신뢰구간이 95%의 신뢰도로 모집단의 파라미터를 포함할 것이라는 것을 의미한다.

(여러 번의 추정 중 **95%가 모집단의 파라미터를 포함**할 것이라는 뜻)

Range(범위) = 최대값 – 최소값

표본이 매우 많으면 median 과 mean이 거의 중복되는데 이러면 대표값으로 써도 충분

분산 = 평균에서부터 얼마나 퍼져있는가?

CV(변동계수) = scale 보정해서 비교하는 것. (자동차와 비행기의 가격차이 scale을 보정)

Coefficient(계수) = X와 Y의 관계성을 설명하는 계수. 이를 추론하는 것이 통계분석의 목표. (날씨와 코로나의 상관관계 설명 등) ///// Y= Ax+B 일 때 x가 바로 계수지

1. Mean – 평균 (평균은 all data를 대표함)
2. Trimmed Mean(절사평균) – 극단값을 제거하고 구한 평균
3. Weighted Mean(가중평균) – 각 데이터에 가중치를 부여, 전체 데이터 합을 가중치 총합으로 나눈 값

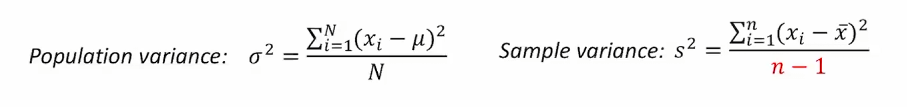
텍스트, 폰트, 화이트, 도표이(가) 표시된 사진

자동 생성된 설명- 특정 값에 가중치를 부여해서 그 값의 중요도를 조절할 수 있음

Bias = 편향 = Sample – True Value

(통계학에서 산포 = 평균에서 얼마나 떨어져 있는가?)

Good sample은 unbiased



**Population variance(모분산) vs Sample variance(표본분산)**

모분산은 N으로 나누어줌 = 분산 과소추정 편향이 있음 = biased variance (편향추정)

표본분산은 n-1으로 나누어줌 = 분산 편향이 없음 = unbiased variance (불편추정)

<자유도 때문에 n-1을 해줘야함>

Box plots 텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 화이트이(가) 표시된 사진

자동 생성된 설명

챕터 1.3 Probability (확률)

Repeatability – 무수히 반복을 했을 때 귀결되는 확률

결합확률(Joint) / 조건부확률(Conditional) / 주변확률(Multiplication)

텍스트, 폰트, 스크린샷, 화이트이(가) 표시된 사진

자동 생성된 설명

**Conditional = A가 일어났을 때 B가 일어날 확률(조건부)**

**Joint = A랑 B가 같이 일어날 확률(결합)**

텍스트, 폰트, 화이트, 대수학이(가) 표시된 사진

자동 생성된 설명

독립이지 않다 = 연관성이 있다 //// 독립이다 = 연관성이 없다

<시험> - A와 B의 바운더리가 안 겹친다는 것이 독립이라는 이야기는 아니다.

스케치, 그림, 원, 화이트이(가) 표시된 사진

자동 생성된 설명 - exclusive event(서로소)

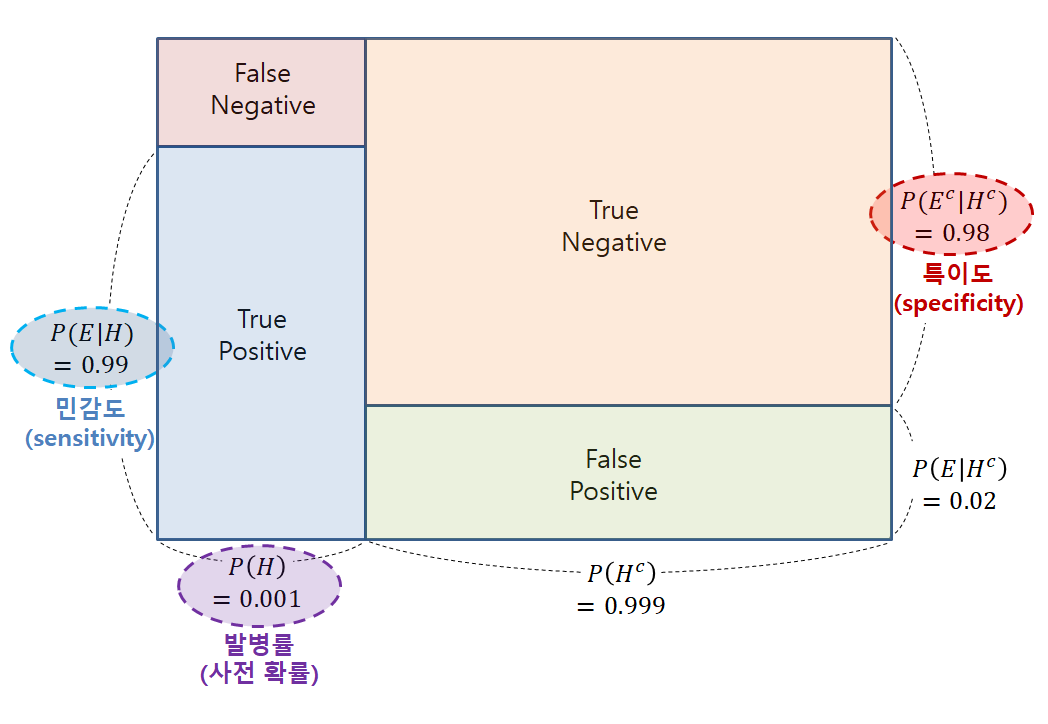
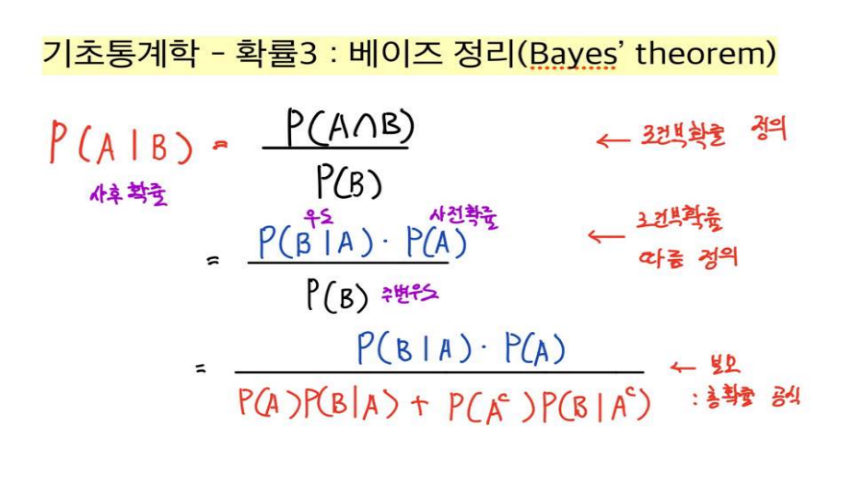
이건 독립이 아님!!! 이건 서로 배타적 연관성을 지니고 있는 것이다.

이건 P(A∩B)=0는 성립하지만 P(A)\*P(B)=0이 아니다!

각각의 사건이 일어날 확률은 존재하기 때문.

<따라서 독립이 성립되려면 자연스럽게 겹치는 부분이 있되, 서로 영향이 없어야 함.>

<베이즈 이론> - 사전확률 = P(A) , 사후확률 = P(A | B)



위양성 위음성은 반비례관계

Sensitivity = 양성도(민감도)

Specificity = 음성도(특이도)

1 – 양성도 = 위양성

1 – 특이성 = 위음성

**베이즈 정리 문제 시험에 나옴**

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

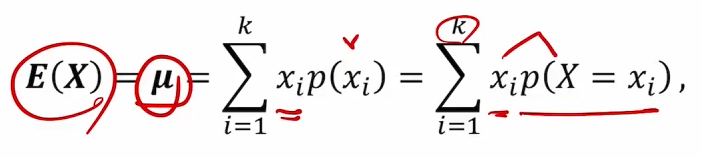
이 계산과정 시험에 나옴.

챕터 1.4 Probability distributions (확률 분포)

**Distribution(분포) 의 종류**

1. 연속형 분포
2. 이산형 분포

E(X) = 기댓값



<시험> - 확률분포의 특징

텍스트, 폰트, 친필, 라인이(가) 표시된 사진

자동 생성된 설명

폰트, 타이포그래피, 서예, 화이트이(가) 표시된 사진

자동 생성된 설명&폰트, 타이포그래피, 친필, 서예이(가) 표시된 사진

자동 생성된 설명&폰트, 타이포그래피, 서예, 화이트이(가) 표시된 사진

자동 생성된 설명

**The Binomial distribution(이항분포)**

* 베르누이 분포 : 1번 시행한 분포
* B(N,P) -> 베르누이(시도횟수, 성공확률)
* 베르누이분포 B(1,p), E(X)=p, Var(X)=pq
* **이항분포 B(n,p), 기댓값 = np, 분산=npq**
* 폰트, 텍스트, 화이트, 서예이(가) 표시된 사진

  자동 생성된 설명

**Normal distribution(정규 분포)**

* 정규분포를 그리기 위해 필요한 values = parameters(평균, 분산)

**Standard Normal distribution(표준 정규 분포)**

* 정규분포를 표준정규분포로 변환하면 유용함
* 표준정규분포에서는 Z-value를 활용한다 -> N(0, 1)

텍스트, 폰트, 도표, 라인이(가) 표시된 사진

자동 생성된 설명시계, 폰트, 라인, 번호이(가) 표시된 사진

자동 생성된 설명(표준화 과정)

**특정 정규분포의 Z-value를 계산하면 N(0,1)인 표준정규분포를 얻을 수 있다.**

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

Sample의 Statistics를 구해서 population의 parameters를 구함 -> population 추정가능

챕터 1.5 Sampling distributions (표본 분포)

시행을 무수히 반복하면 우린 sample statistics으로부터 sample의 Distribution를 얻을 수 있음.

친필, 폰트, 서예, 디자인이(가) 표시된 사진

자동 생성된 설명<- 중심극한정리

모집단을 무수히 확보했다고 해도 그 모집단평균이 정규분포라고 확신할 수 없음.

그러나 표본집단을 무수히 확보했다면 표본평균의 분포는 반드시 정규분포를 따름. (CLT)

<시험에 나옴!>

모분산과 표본분산 구하는 법의 차이

시그마텍스트, 폰트, 영수증, 라인이(가) 표시된 사진

자동 생성된 설명 제곱 = Population Variance

S제곱 = Sample Variance

**Distribution of the ‘sample mean’ (표본 평균의 분포) X bar = sample mean**

1. 모집단에서 표본을 뽑아 평균을 냄 = 표본평균
2. 표본평균을 여러 번 추출하여 평균을 냄 = 표본평균의 평균 = 모집단의 평균

즉 

1. 표본평균의 분산은 모집단 분산을 표본크기 n으로 나눈 값. 따라서 n이 커지면 표본평균의 분산은 감소함 폰트, 스크린샷, 라인, 화이트이(가) 표시된 사진

   자동 생성된 설명

폰트, 텍스트, 친필, 화이트이(가) 표시된 사진

자동 생성된 설명

이게 매우 중요한 부분인데,

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

표본들을 기술통계를 활용하여 모집단의 파라미터를 구할 수 있음(뮤, 시그마^2)

그러면 그 표본들은 모집단의 정규분포 N(뮤, 시그마^2)을 따르겠지.

텍스트, 폰트, 화이트, 스크린샷이(가) 표시된 사진

자동 생성된 설명근데 표본평균의 정규분포는 N(뮤, 시그마^2/n)을 따른다고 ㅇㅋ?

차이점. 표본(X)의 분산=시그마^2, 표본평균(X바)의 분산=시그마^2/n

CLT(중심극한정리) – 표본크기가 매우 크면 표본평균의 분포가 정규분포에 수렴함.

1. 모집단이 정규분포를 따르지 않아도 됨. (모든 형태의 분포에 적용가능)
2. 모집단의 평균과 분산은 존재해야 한다.

텍스트, 폰트, 스크린샷, 화이트이(가) 표시된 사진

자동 생성된 설명

이거 풀어보기. 주의점 = 표본에 대하여가 아니고 표본평균에 대하여 구하는거임.

~~Z-test, T-test 한번 읽어보기~~?

이 수업의 목표는 2개의 그룹의 평균을 비교하는 것.

텍스트, 스크린샷, 폰트, 대수학이(가) 표시된 사진

자동 생성된 설명

폰트, 도표, 라인, 디자인이(가) 표시된 사진

자동 생성된 설명이때의 Z값 공식.

Z-test는 표본의 숫자가 클 때 사용함

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

챕터 1.6 Estimation(추론) – 매우중요 완벽하게 숙지!

**Statistics Estimation(통계적 추론)**

1. Estimation – 추정 (모집단의 parameter 구하기)
2. Hypothesis Testing – 가설검정. (가설이 참인가 거짓인가?) / p-value구하기

**Estimation**

1. Point estimation(점 추정)

* 모집단을 추정하기 위해 하나의 숫자값(mean, median, variance…등등) 사용

1. Interval estimation(구간 추정)

* 모집단의 모수를 추정하기 위해 2개의 숫자값을 사용.
* 2개의 숫자값을 이용하여 모집단의 모수가 그 구간 사이에 있을 확률(신뢰구간)을 계산함.

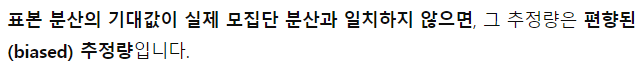
텍스트, 스크린샷, 폰트, 대수학이(가) 표시된 사진

자동 생성된 설명

폰트, 텍스트, 화이트, 서예이(가) 표시된 사진

자동 생성된 설명추정한 Parameter가(ex. Mean median… 등) 이 정확하다면(편향이 없다) E(추정량) = 실제의 parameter

따라서 unbiasedness(불편향) parameter



Confidence interval for population mean = Uncertainty of our estimation

신뢰구간 for 모집단 평균 = 우리 추정의 불확실성 반영

**표본 100, 평균85, 표준편차10 일때**

텍스트, 폰트, 친필, 화이트이(가) 표시된 사진

자동 생성된 설명

<Z검정> - 표본이 30개 이상일 때

스크린샷, 라인, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명구간추정시 Z값 구할 때 모표준편차를 알면 이거쓰면되고

폰트, 도표, 라인, 번호이(가) 표시된 사진

자동 생성된 설명

모르면 모표준편차 대신에 표본표준편차 쓰자

<T검정> - 표본이 30개 이하일 때

라인, 스크린샷, 도표, 디자인이(가) 표시된 사진

자동 생성된 설명.

그냥 모표준편차를 모르면 무조건 이 식을 사용하는데.

표본이 30개 이상이면 Z검정

30개 이하면 T검정 하면 됨

텍스트, 폰트, 스크린샷, 화이트이(가) 표시된 사진

자동 생성된 설명

이거 풀어보기(Z검정)

텍스트, 폰트, 스크린샷, 화이트이(가) 표시된 사진

자동 생성된 설명

이거 풀어보기(T검정)

친필, 텍스트, 폰트, 화이트이(가) 표시된 사진

자동 생성된 설명

두 집단의 평균 차이를 비교할 때

그 평균 차이의 존재구간을 정함.

보통 Z=1.96 (95%신뢰구간)

텍스트, 폰트, 스크린샷, 대수학이(가) 표시된 사진

자동 생성된 설명

2개의 표본집단에 대하여 평균 차이를 구간추정했을 때

그 범위가 0을 포함하지 않는다면 2개의 표본집단의 평균은 절대 같을 수 없음.

**우리의 sample이 normal distribution이면**

1. 샘플이 30개 이상인가? -> 모집단의 분산을 안다 - Z검정

모집단의 분산을 모른다 – Z or T

1. 샘플이 30개 미만인가? -> 모집단의 분산을 안다 – Z

모집단의 분산을 모른다 -> T검정

**우리의 sample이 not normal distribution이면**

1. 샘플이 30개 이상인가? -> Z
2. 샘플이 30개 미만인가? -> 못구함 ㅋㅋ

따라서 모분산을 모르는 경우 T를 쓰면 됨.

나머지의 경우 죄다 Z쓰면 됨

<In General 두개의 모집단의 모평균을 비교할 때> - 모분산을 모를 때(T-test)

1. 두 모분산이 같다면? -> Pooled T-test
2. 두 모분산이 다르다면 -> Welch’s T-test

텍스트, 스크린샷, 폰트, 대수학이(가) 표시된 사진

자동 생성된 설명

**모집단 비율의 신뢰구간 추정**

폰트, 화이트, 번호, 상징이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷, 화이트이(가) 표시된 사진

자동 생성된 설명

풀어보기

**두 모집단 간의 모집단 비율 차이를 검사. 통계적으로 유의미한지 판단**.

텍스트, 폰트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷, 영수증이(가) 표시된 사진

자동 생성된 설명

챕터 1.7 Hypothesis testing(가설 검정)

가설검정의 목표 : make a decision

대립가설이 옳다는 주장의 **통계적 증거**를 찾는 과정

Null Hypo = 귀무가설 Alternative Hypo = 대립가설

귀무가설을 기각할 수 없다는 것이 귀무가설이 옳다는건 아니다.

**귀무가설을 기각할 증거가 부족하다는 것이 귀무가설이 참임을 의미하지는 않습니다**. 단지, 데이터를 바탕으로 귀무가설을 **기각할 만큼의 통계적 증거**가 충분하지 않았다는 것뿐입니다.

1. **제1종 오류 (Type I Error): 귀무가 맞는데 대립을 채택**

- 귀무가설이 참인데도 불구하고, 귀무가설을 기각하는 오류입니다.

(내말이 맞다고!! 시발 맞다니까? -> 대립가설 채택)

2. **제2종 오류 (Type II Error):** **대립이 맞는데 귀무를 채택**

- 귀무가설이 거짓인데도 불구하고, 귀무가설을 채택하는 오류입니다.

(내 말이 맞나? 아닌것 같은데? -> 그럼 그냥 귀무가설 채택)

제1종 오류와 제2종 오류는 서로 반비례 관계에 있습니다.

보통 1종 오류가 더 중요합니다. 2종 오류를 범해도 현상유지. 1종 오류를 범하면? ㅈ됨

**POWER(검정력)** = 1 – Type2 Error(참을 버리고 거짓을 택함)

= 대립가설이 참일 때 맞다고 할 경우. (True일 때 positive)

따라서 POWER가 0.8 이라면 대립가설이 참일 때 대립가설을 채택할 확률이 80%

이 경우 Type2 Error는 0.2

친필, 폰트, 서예, 타이포그래피이(가) 표시된 사진

자동 생성된 설명Z나 T value가 올라가면 P-value는 감소함

Z나 T는 test값이고 P는 result value

**Confidence Interval(신뢰구간)**

**텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명**

**P-value**

p-값은 **귀무가설이 맞다고 가정**한 상황에서 관측한 데이터를 얻을 확률.

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

P값은 Z나 T를 구한다음 표준정규분포표 테이블에서 수치를 구한다음 1에서 빼주면 됨.

**즉 Z or T value의 표준정규분포표 테이블수치 + P value = 1**

Confidence Interval이나 P값이나 둘 다 **통계적 불확실성**을 설명함 (정비례 관계)

1. **두 표본평균을 비교할 때. Confidence Interval이 0을 포함하면 P>0.05, Z<1.96**

* 신뢰구간이 0을 포함하면 두 표본평균이 같을수도 있잖아?
* 그러면 대립가설이 귀무가설을 대체해서는 안되는거야. 같은 것일 수 있으니까.

1. **Confidence Interval이 0을 포함하지 못하면 P<=0.05, Z>1.96**

* 신뢰구간이 0을 포함하지 못하면 두 표본평균은 절대 같을 수 없지.
* 그러면 대립가설이 귀무가설을 대체할 수 있네?

텍스트, 폰트, 스크린샷, 화이트이(가) 표시된 사진

자동 생성된 설명

(책 136페이지부터 154까지 예제 쭉 나오는거 3개)풀어보기 123 456 789

**챕터3 Linear Regression analysis(선형 회귀 분석)**

Sensitivity = 민감도(실제론 True인데 -> 테스트 Positive일 확률) 양성/유병

Specificity = 특이도(실제론 False인데 -> 테스트 Negative일 확률) 음성/무병

<헷갈리는 경우>

Positive predictive value = 양성예측도 (테스트 양성 -> 실제 True일 확률)

* 테스트Positive / All Positive (유병/양성)

Negative predictive value = 음성예측도 (테스트 음성 -> 실제 False일 확률)

* 테스트Negative / All Negative (무병/음성)

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

이 실험은

민감도, 특이도를 위한 설계

보통 모든 모델들은 예측을 하기 때문에 양성예측도(PPV)가 중요함.

내 모델이 양성을 띄웠는데 그게 진짜 True냐? 인지가 궁금한거거든.

폰트, 텍스트, 화이트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명g(X)의 기댓값 = 확률변수\*확률 의 총합

**챕터3.1 Regression Data(회귀 데이터)**

X로 Y예측하기. X와 Y의 연관성 설명. (상관관계를 설명하는 것임. 인과관계가 아니고)

Y = dependent variable (종속 변수) - Outcome

X = independent variable (독립 변수) – Explanatory variable

**챕터3.2 Simple Linear Regression**

텍스트, 폰트, 타이포그래피, 화이트이(가) 표시된 사진

자동 생성된 설명폰트, 텍스트, 타이포그래피, 디자인이(가) 표시된 사진

자동 생성된 설명



텍스트, 폰트, 스크린샷, 대수학이(가) 표시된 사진

자동 생성된 설명

**위 3개의 가정을 만족해야 회귀분석이 작동할 수 있음.**

**도표, 라인, 그래프, 텍스트이(가) 표시된 사진

자동 생성된 설명**뮤만 다름

이게 위 3개의 가정을 만족하는 경우라는 거지. (이상적인 선형회귀)

텍스트, 폰트, 라인, 도표이(가) 표시된 사진

자동 생성된 설명 Residual = Error

Error가 정규분포를 따르기 때문에 결국 도 정규분포를 따른다.

* 가 정규분포를 따르지 않으면 회귀분석에 사용할 수 없음.

따라서 가 정규분포를 따르게끔 여러 변환을 사용할 수 있음

<시험>

폰트, 타이포그래피, 서예, 친필이(가) 표시된 사진

자동 생성된 설명, 

B1에 대하여 설명하라.

B1은 X와 Y평균에 대한 상관관계를 보여주는 계수입니다.

(X가 변할 때 Y의 평균이 얼만큼 이동하는가)

Parameter Estimation(LSE) – Least Squares Error

가장 좋은 선형회귀모델은 뭐지? -> 모든 에러의 총합이 제일 작을 때!

에러는 이렇게 구함

그럼 B0, B1을 구하는 법은? -> 오차 제곱 합 최소화

텍스트, 폰트, 화이트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

<시험>

폰트, 친필, 화이트, 서예이(가) 표시된 사진

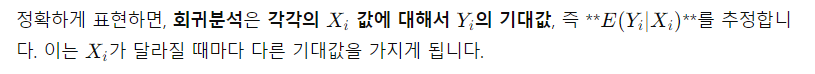
자동 생성된 설명친필, 폰트, 텍스트, 라인이(가) 표시된 사진

자동 생성된 설명

본 수업의 목표는 회귀분석을 통해서 X와 Y의 관계성을 설명하는 것임.

X로 Y를 정확하게 예측하는 것은 목표가 아님. 오히려 이것은 머신러닝의 영역

따라서 relationship을 목표로 할 때 predict 성능으로 평가하는 건 맞지 않음



선형회귀를 통해