**평균**

ㅇ 대푯값 중 1순위

데이터 셋의 대푯값을 나타낼 때, 사용되는 방법이다.

ㅇ 중위값, 빈도값

**분산**

데이터 셋의 대표값은 알지만

데이터 셋을 이루는 각 관측치의 값은 모르기 때문에

대표값(평균)으로부터 떨어진 정도로 개별 데이터의 분포를 알기 위함이다.

마찬가지로 전체 편차로부터의 평균으로 대푯값을 통해 분산을 표현한다.

* 대푯값을 기준으로 개별 값의 분포를 나타내기 위함

#제곱을 하여 구하기 때문에 값이 커지고, 단위가 바뀌는 특징이 있다.

**표준편차**

분산의 제곱근

구하고자 했던 편차의 절대값으로 볼 수 있다.

**검정**

평균 : 중심극한의 법칙을 따른다.

즉) 데이터 셋에서 임의로 뽑았을 때 그 값은 평균에 가깝다 = 정규분포를 띈다.

> 정규분포의 특징 : 중심에서 값이 많이 분포하고 극단으로 갈수록 그 분포가 줄어드는 데이터의 분포성

즉) 극단으로 갈수록 잘 뽑히지 않는다.(이상치)

> 이 특징 (대푯값 평균에서부터 떨어진 극단으로 갈수록 이상치 값이다 = 우연이다/우연이 아니다)을

갖고 검정이라는 것을 사용한다.

* 검정의 근거 : 신뢰구간, 유의수준(관례 95%)을 갖고 데이터 셋의 이상치를 판별

**T-분포**

정규분포는 어느 정도 개수가 보장되어야 한다.

그러나 T분포를 통해 개수가 적더라도 정규분포와 똑같지는 않지만 정규성의 곡선을 그리는 것을 확인시킨 분포

(\*비교하는 방법은 표준정규분포화 시켜서 비교하면 된다.)

T분포의 공식과 자유도(개수 : t 분포의 파라미터)에 따른 값의 수치를 다 계산해 놓았음

(자유도가 무한이면 정규분포와 같다.)

즉\_ 개수가 적을 때 띄는 정규성을 스튜던트라는 사람이 표준화 시켜놓은 것이고,

이것을 통해 개수가 적은 데이터 셋에도 검정과 추정을 사용할 수 있다.

ㄴ

t-test

모집단의 표준편차가 알려지지 않았을 때, 정규분포의 모집단에서 모은 샘플(표본)의

평균값에 대한 가설검정 방법

목적 : 두개의 집단이 같은지 다른지 비교하기 위해 사용한다.

집단 : 통계에서 일반적으로 집단이란 샘플(표본)을 이야기한다.

모집단도 이야기할 때 있음(모집단, population)

모집단 vs 표본(샘플)

N / n

뮤 / 엑스바

시그마제곱 s제곱

시그마 s

어떻게 두 집단이 같은지 다른지 비교할까

: 두 집단의 평균값이 통계적으로 같은지 다른지를 확인

평균값이 데이터 셋의 대표값이기 때문.

Q. 178.5 vs 179.9

통계적 질문

1.A집단의 평균키와 B집단의 평균키가 우연히 같은 확률은 어떻게 될까?

2. = A집단과 B집단의 평균키 차이인 1.4차이가 우연히 발생했을 확률은 얼마나 될까?

A

* 우연히 발생 : 두 집단의 키는 통계적으로 차이가 없는 것이고(1.4cm차이는 의미가 없다.)
* 우연히 발생하지 않았다면 : 두 집단의 키 차이는 통계적으로 의미가 있는 것이다.

**\*우연히 발생한다 = “평균적으로 자주 나타난다.(평균에 위치한다)”**

**\*우연히 발생하지 않는다 = “평균적으로 나타나지 않는 값이다.”**

>> 판단근거

1.4cm의 차이는 과연 우연히 발생하지 않았을 만큼 큰 것인가

우연히 발생할 정도로 작은 것 인가?

그렇다면 얼마나 커야 우연이 아니고, 얼마나 작아야 우연일까?

즉, 1.4cm이라는 두 집단의 차이만 보았을 때는 과연 의미가 있을 정도만큼 차이가 큰지

(1.4CM이 얼마나 큰지 혹은 작은지) 알 수 없다.!!

* 우리는 이제 이 1.4CM이 얼마나 큰지 혹은 작은지 결정할 나름의 비교대상이 필요하다

(우연인지, 우연이 아닌지에 대해 판단을 해줄 기준대상).

* 그렇다면 누구를 가지고 와서 비교해야 할까
* 표준편차(분산) 밖에 없다.

: 왜 표준편차가 비교의 대상일까?

EX. {1,2,3,4,5}이 있을 때 평균 : 3 , 분산 :2.5 표준편차 약 1.58

이것의 의미 : 평균 값 3을 중심으로 평균적으로 관측치들이 약 1.58정도 떨어져 있다.

퍼짐의 정도; -1.58+3 & +1.58+3 의 범위에서 퍼져있다.

그렇다면!!!

* +- 1.58정도 펴져 있다는 것은 의미 있는 중요한 퍼짐일까? 아니면 우연히 발생한 퍼짐일까?
* 답 : 그렇다, 표준편차는 데이터에 큰 문제가 없는 한 의미가 없는!! 즉 우연히 퍼져 있는 정도이다.
* 다시 말하면, 우리의 데이터는 평균값 3을 중심으로 랜덤하게!!(보통 이렇게 말한다) 1.58정도씩 좌우로 퍼져있는 것이다. 어떠한 규칙성 없이 1.58정도씩 우연히 퍼져있는 것

다시 1.4CM으로 돌아가면..

이 1.4CM의 차이도 결국 두 집단의 평균적인 거리입니다.

: 왜냐하면, 두 집단의 수많은 데이터의 평균의 차이가 1.4CM라는 것은

두 집단의 수 많은 데이터 사이의 평균적인 거리가 1.4CM 이라는 것이기 때문이다.

* 두 집단이 1.4CM정도 떨어져 있다.(퍼져있다.)

그렇다면!! 비교해보자

1. 두 집단 A와 B의 데이터 사이의 평균적인 거리는 1.4CM이다.
2. 두 집단 A와 B의 데이터들의 표준 편차가 XXXcm 이라면!!?

따라서

1. 만약 1.4cm가 표준편차 XXXcm보다 현저히 작다면, 우리는 이 1.4cm의 차이에 큰 의미를 둘 수 없다.
2. 그러나, 1.4츠가 표준편차 XXX츠보다 현저히 크다면, 우리는 이 1.4cm의 차이에 큰 의미를 둘 수 있을 것입니다.

결론적으로

* **두 집단의 평균값의 차이가 두 집단의 표준편차보다 현저히 작으면, 우리는 이 차이가 우연히 발생했다라고 결론을 내릴 것 이다. (우연히 발생했다 = 평균적으로 있는 볼 수 있는 일이다.)**
* **반대로, 두 집단의 평균값 차이가 표준편차보다 현저히 크면, 우리는 이 차이가 우연히 발생하지 않았다고 결론을 내릴 것입니다. (우연히 발생하지 않았다 = 이상치이다.)**
* **통계학이란 분산(표준편차)의 마법, 이제 그 마법의 시작된다.**

이제,

1. 여기서 어떻게 두 집단의 표준편차를 구할지

2. 또, 현저히 크다/작다를 어떻게 결정할지에 대해서 알아보자.

T-test를 바로 시작하지 못하는 이유

1. 정규분포

: t-test를 이해하기 위해서는 정규분포로 몸을 풀고 가야 한다.

: 대부분의 경우, 정규분포의 z-test와 t-test를 단순/별도로 설명하지만 연결해서 볼 필요가 있다.

: 본질적으로 같은 것이기 때문이다.

: 그래서 여기서 정규분포를 이해하고 넘어가야 한다.

정규분포 – 가우스가 만듬

[특징]

-종모양

-정가운데(평균)을 중심으로 좌우대칭

-정규분포의 양 끝은 영원히 0에 닿지 않음

[정규분포는 평균과 표준편차만으로 규정됨]

-평균과 표준편차가 다른 무한대 개의 서로 다른 정규분포가 존재

[정규분포의 아래의 면적은 **확률**을 의미함]

> 정규분포 곡선은 그래서 확률곡선이라고 이해해도 된다.

-정규분포 곡선 아래의 모든 면적의 합은 1

-따라서 정규분포를 이용한 확률은 적분을 통해서 구해야 한다.(미친짓)

* 왜 미친짓인가.. 확률을 구할려면 적분을 해야하는데

: 표준정규분포!!!를 학자들이 만들어놓았다.

표준정규분포란,

평균=0, 표준편차가 1인 정규분포

왜 만들었을까?

: 무한대 가지의 정규분포 곡선을 적분하는 번거로움을 덜기 위해

즉, 표준정규분포로 기존의 어떤 정규분포라도 다 바꾸어서 적분하지 않고 쉽게 확률을 구하게 하기 위해서

표준정규분포 예시

1000명 대상 영어 시험, 영어점수의 분포가 정규분포에 근사 / 평균 82 , 표준편차 5

이때, 82점부터 90점까지의 점수를 받은 학생의 수는?

82~90점까지의 아래 면적을 적분을 적용해 구하면, 이 면적은 곧 확률이므로

구한 확률 \*1000명을 적용하면, 82점~92까지의 점수를 받은 학생의 수를 구할 수 있다.

(근데 이걸 적분할 수 있을까???..)

1. 표준 정규분포로 바꾼다. (z-score)
2. 통계학자들이 표준정규분포표라는 확률표를 만들어서, 모든 z-score에 대한 확률을 적어놓았음.
3. 따라서 90의 z-score를 확률표에서 찾아 확률을 구하면 된다.
4. (0.9452-0.5)\*1000

* 이와 같이 z-score(z값)을 가지고 하는 테스트를 z-test라고 한다.
* Z-test는 z값과 표준정규분포를 이용하여 할 수 있다.

z-socre(z값)으로 변환하는 것은 z-transformation이라고 하기도 하고,

표준화(standardization)이라고 하기도 한다.

: 1의 표준편차당 관찰값(x)가 평균으로부터 얼마나 떨어져 있는지를 의미

: z값은 단위로부터 자유롭다.

이제 정리해보자..

정규분포곡선의 아랫쪽 면적이 확률이라고 했다.

이게 어디로 연결이 될까?

* 첫 강의에서 이야기 했던 내용으로 가보자
* **통계적인 사고란**, 어떤 사건이 **우연**히 발생할 확률이 얼마일까?
* **즉, 이 사건이 얼마나 평균적으로 발생하는 사건인가?**
* 여기서 말하는 확률이 바로 정규분포곡선 아랫쪽의 면적인 그 확률입니다.!!

이것을 우리가 처음 하려고 했던 t-test의 질문을 떠올려보면,

* **A대학 남학생 평균키(178.5cm)와 B대학 남학생 평균키(179.9cm)가 우연히 같은 확률은 얼마나 될까?**

**(= A대학 대표값인 평균 키와 B대학의 대푯값 평균키와의 차이는 평균적으로 있을 수 있는 일인가?)**

* 다만, t-test를 할 때는 정규분포를 쓰지 않고 t분포 곡선을 사용하나, 개념적으로는 같다.

t-test를 바로 시작하지 못하는 이유

1. 양측검정 vs 단측검정

t- test를 이해하기 위한 다음은 양측/단측검정을 이해하는 것 입니다.

앞에 나왔던 t-test 예를 통계적 검정에 연결해 보자.

< 양측검정 >:

귀무가설 : A집단 키 평균 = B집단의 키 평균

대립가설 : A 집단 키 평균 != B집단 키 평균 ( A>B or A<B)

귀무가설 : A-B =0

대립가설 : A-B >0 , A-B<0

< 단측검정 >

귀무가설 : A =B

대립가설 : A>B \_ 하나만 설정 (A<B 로 하나로 설정해도 된다)

언제 양측검정을 하고, 언제 단측검정을 하는가?

* 이것은 정해진 바가 없다.
* 연구자가 단순하게 다를 것이다 라고 판단하는 경우는 양측검정
* 이전 연구나, 이론 등의 충분한 근거를 통한 확실한 판단이 있는 경우 단측검정

A와 B대학의 남학생 평균키 차이 1.4cm가 우연히 발생했을 확률은 얼마나 될까?

**우연일 경우 : 0(평균) 근처 위치한다면 우연**

**우연이 아닐 경우 : 극단 위치한다면 우연이 아니다**

양측검정은 우연이 아닐 경우 0.25

단측검정은 우연이 아닐 경우 0.5

결론적으로 양측검정과 단측검정의 차이는 대립가설

* 양측은 0보다 크거나 작은 두가지 모두 포함하므로, 분포곡선의 양쪽 꼬리의 면적의 합이 5%(2.5+2.5)에 들어갈 만큼 크거나 작아야 한다.
* 단측은 0보다 크다(우측검정)와 0보다 작다(좌측검정)의 두 가지로 나누어볼 수 있고 어느 한쪽 꼬리의 면적이 5%에 들어갈 만큼 크거나 작아야 한다.
* 0을 기준으로 양측이든 단측이든 95%안에 들어오면 두 평균값의 차이인 Da-b는 우연히 발생한 것이므로 두 집단의 평균값은 통계적으로는 같은 것이다.

이제 T-test를 해보자

H0 : A대학 평균178.5 = B대학 평균179.9

H1 : A대학 평균 != B대학 평균

우리는 양측검정

A와 B의 차이가 있는 것 같은데 차이가 있어야 하는 이유나 원인을 모르겠어서, 일단 다를 것이다

z-test를 떠올려 보자

( z-test는 표준화작업을 통한 비교 )

z-test를 하기 위해 필요했던 것은 : z-값(z-value)와 표준정규분포를 이용해서 확률값을 구했다.

따라서 우리는 이제 t-값과 t-분포가 필요하다.

다시 한번 더, t-test의 목적은 무엇이었나

: 두 집단이 같은지 다른지 알고 싶을 때!

T-value = A평균 -B평균 / 표준오차 : (표본의 표준편차 / 표본의 개수의 루트)

우리의 목적

두 집단의 평균값이 같은지 다른지 알고 싶다.

그래서 통계적 가설에 의거하여, 두 집단의 평균값의 차이가 0과 같은지 궁금합니다.

위의 값에서 우리가 궁금해 하는 것은 그 차이는! 분자에 있습니다.

여기서부터 중요한 것이 통계적인 생각/질문/접근법입니다.

도대체 저 값이 얼마나 커야 큰 것일까?

대소를 알기 위해서는 비교대상이 필요하다.

우리는 이 두 평균값의 차이(분자)를 표준편차와 비교하는 겁니다.

왜?

: 표준편차란 우리의 데이터가 평균값을 기준으로 평균적으로 퍼진 정도입니다.

: 따라서 이 자체는 의미 없는 편차

: 만약 두 집단의 평균값의 (편)차가 의미 없는 편차인 표준편차만도 못하다면, 당연히 이 차이는 우연히 발생했다고 보아야 할 것입니다.

: 그런데 여기서 루트n의 역할은 무엇일까