

마케터를 위한 기초통계

- 디지털 마케팅 SCHOOL

수업의 흐름

- 큰 틀 그리기

Chapter 2. 통계의 언어 배우기

- 확률변수
- 확률분포
 - 정규분포
 - 표준정규분포

2.8. 확률변수 (Random variable)

- 확률변수란 일정한 확률을 가지고 발생하는 사건에 수치를 부여한 것
- 보통 X 로 표시함
- 동전 한 개를 던질 때, 모든 가능한 사건의 집합은?
 - $S = \{ \text{앞면}, \text{뒷면} \}$
- 그러나 실제 통계학적 방법 및 분석과정에 들어서면, 어떠한 수치를 부여할 필요가 있다
 - $S = \{ 1, 0 \}$: 앞면에는 1, 뒷면에는 0을 부여

2.9. 확률분포 (Probability Distribution)

- 동전을 두 번 던질 때, 모든 가능한 사건들과 각 사건이 나타날 확률에 대해서 다음과 같이 정의 가능하다
- 모든 가능한 사건의 집합 : $S = \{HH, HT, TH, TT\}$

사건	앞면의 수	각 사건의 확률
{ H , H }	2	1/4
{ H , T }	1	1/4
{ T , H }	1	1/4
{ T , T }	0	1/4

- 만약 앞면이 나올 횟수를 확률변수(X)로 하고자 한다면, 다음과 같이 확률분포를 정의 가능하다
- 표본공간 $S = \{2, 1, 0\}$

확률변수(X_i)	$P(X_i)$
0	1/4
1	1/2
2	1/4

- 즉, 확률분포란 어떤 확률변수가 취할 수 있는 모든 값들과 이 값들이 나타날 확률을 표시한 것이다.

확률분포 (예제) - 문제

- 동전 3개를 던질 때 앞면이 나오는 사건에 대한 확률분포표와 확률분포를 그려보기

2.9.1. 확률분포의 기대값 (Expected Value)

- 확률분포의 평균값 (average, weighed average)
- 표기법 : $E(X)$
- 기댓값의 계산

$$E(X) = \sum X_i \cdot P(X_i)$$

2.9.1. 기대값의 특성

1. 확률변수 X 에 일정한 상수 a 를 곱한 확률변수의 기댓값은 확률변수 X 의 기댓값에 a 를 곱한 것과 같다.

$$E(aX) = a \cdot E(X)$$

2. 확률변수 X 에 일정한 상수 b 만큼을 가감한 확률변수의 기댓값은 확률변수 X 의 기댓값에 b 를 가감한 것과 같다.

$$E(X \pm b) = E(X) \pm b$$

3. 위의 두 가지 결과를 결합하면 다음 식이 성립된다.

$$E(aX \pm b) = a \cdot E(X) \pm b$$

확률분포 (예제) - 문제

- 위에서 본 3개의 동전던지기 문제를 활용하여 동전의 앞면이 나올 기댓값을 구하세요.

- 만약 동전의 앞면의 갯수 당 500원을 받는다고 한다면, 동전 3개 던지기의 기대 수익은 얼마인가?

2.9.2. 확률분포의 분산 (Variance)

- 확률분포의 분산
- 표기법 : $Var(X)$
- 분산의 계산

$$\begin{aligned} Var(X) &= \sum [X_i - E(X)]^2 \cdot P(X_i) \\ &= E[\{X - E(X)\}^2] \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

2.9.2. 분산과 표준편차의 특성

1. 어떤 확률변수에 일정한 상수를 더한 확률변수의 분산은 본래의 확률변수의 분산과 같다. 확률변수에 상수를 더하는 것은 분포의 분산도에는 아무런 영향을 미치지 못하기 때문이다.

$$Var(X + b) = Var(X)$$

2. 어떤 확률변수에 일정한 상수 a 를 곱한 확률변수의 분산은 본래의 확률변수의 분산에 a^2 를 곱한 것과 같다.

$$Var(aX) = a^2 Var(X)$$

3. 위의 두 식을 종합하면 다음과 같은 식이 성립된다.

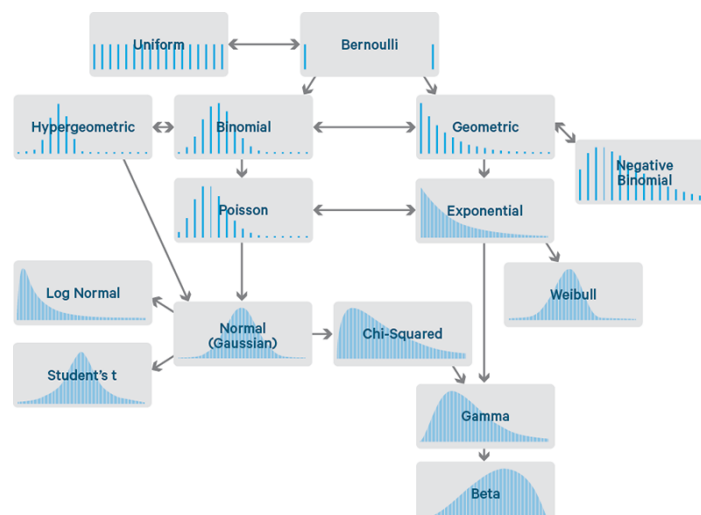
$$Var(aX + b) = a^2 Var(X)$$

확률분포의 기대값 (실습) : 일일 세션과 확률변수의 관계?

- 사건 :
- 확률변수 :
 - ex.
- 확률 :

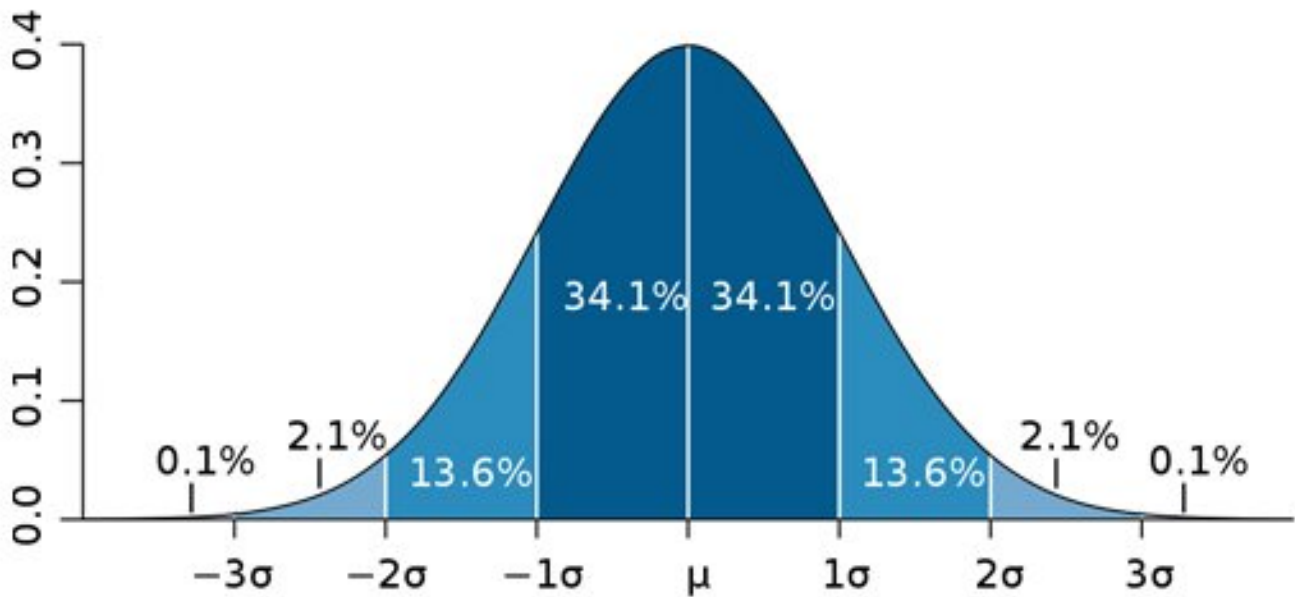
2.9.3. 여러가지 분포 (참고)

- 이항분포 (Binomial)
- 다항분포 (Multinomial)
- 정규분포 (Normal)
- 표준정규분포 (Standard Normal)
- t -분포 (Student's t)
- χ^2 -분포 (Chi-Squared)
- F -분포 (F)

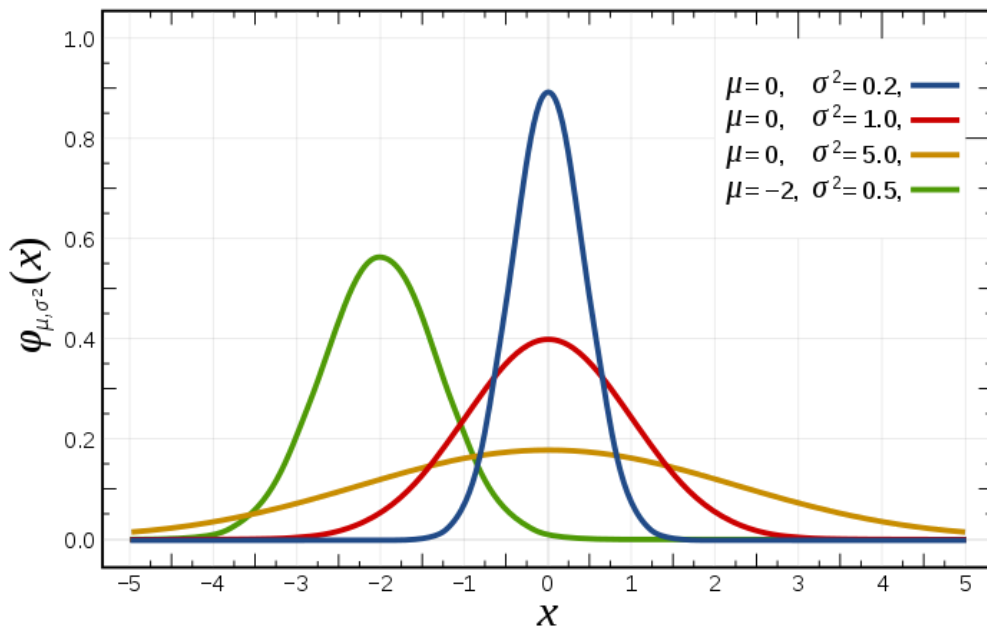


2.9.4. 정규분포 (Normal Distribution)

- 정규분포는 표본을 통한 통계적 추정 및 가설검정이론의 기본
- 실제로 사회적, 자연적 현상에서 접하는 여러 자료들의 분포도 정규분포와 비슷한 형태를 띠
- 현실적인 자료가 이론적인 정규분포와 완전히 일치하는 것은 아니지만 정규분포의 형태에 가깝게 나타나므로 이를 자료분석에 이용할 수 있다는 점



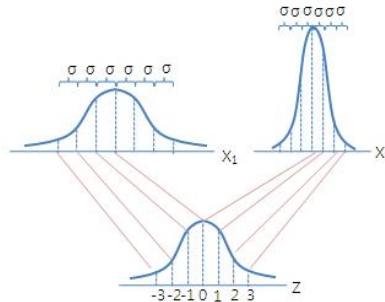
정규분포의 특성



1. 정규분포의 모양과 위치는 분포의 평균과 표준편차로 결정된다.
2. 정규분포의 확률밀도함수는 평균(μ)을 중심으로 대칭인 종모양이다.
3. 정규곡선은 X 축에 맞닿지 않으므로 확률변수 X 가 취할 수 있는 값의 범위는 $-\infty \leq X \leq +\infty$ 이다.
4. 분포의 평균(μ)과 표준편차(σ)가 어떤 값을 갖더라도, 정규곡선과 X 축 사이의 전체 면적은 1이다. (면적 = 확률)

2.9.5. 표준정규분포 (Z-분포)

- 정규분포는 평균과 표준편차에 따라 모양과 위치가 각기 다르기 때문에 두 분포의 성격을 비교하거나 특정 정규분포에서 확률을 계산하기 위해서는,
- 먼저 모든 정규분포의 평균과 표준편차를 표준화하여 표준적인 정규분포(standard normal distribution)를 만들어야 한다



Quiz

- 어느 학생이 영어와 수학 시험을 치렀다. 그 결과 영어점수는 80점이고 수학점수는 75점이었 다. 이 학생은 어느과목을 더 잘했다고 할 수 있는가?
- 추가정보 : 영어과목에서는 전체학급의 평균은 90점, 표준편차는 5점 그리고 수학과목에서는 평균이 60점, 표준편차가 10 점이라고 한다.

표준정규분포 (계속)

- 표준정규분포는 모든 정규분포를 평균 $\mu = 0$, 표준편차 $\sigma = 1$ 이 되도록 표준화한 것이다.
- 어떤 확률변수 X 의 관찰값이 그 분포의 평균으로부터 표준편차의 몇 배 정도나 떨어져 있는가를 다음과 같이 표준화된 확률변수 Z 로 나타내기 때문에 표준정규분포를 Z-분포 라고도 한다.

$$Z = \frac{X - \mu}{\sigma}$$

Quiz (그림도 함께 그려보세요)

- 앞에서 든 예를 Z 의 척도로 바꾸어 보면,

$$Z = \frac{X - \mu}{\sigma}$$

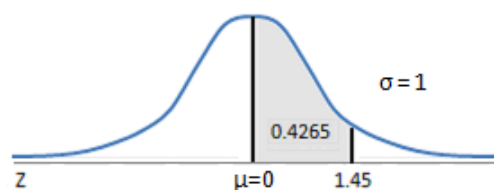
영어 $Z =$, 수학 $Z =$

2.9.6. 표준정규분포표 보는 법

Areas Under the One-Tailed Standard Normal Curve

This table provides the area between the mean and some Z score.

For example, when Z score = 1.45
the area = 0.4265.

[illegible]

정규분포의 확률계산 (예제 1) - 문제

- $Z = 0$ 부터 $Z = 1.5$ 사이에 확률변수가 있을 확률
- $Z = -1$ 부터 $Z = 1$ 사이에 확률변수가 있을 확률
- $Z = -1.5$ 부터 $Z = -0.5$ 사이에 확률변수가 있을 확률
- $Z = -2$ 보다 작거나 $Z = 2$ 보다 큰 사이에 확률변수가 있을 확률

정규분포의 확률계산 (예제 2) - 문제

- 한 초등학교 전교생의 IQ를 측정해 본 결과 평균 $\mu = 100$, 표준편차 $\sigma = 10$ 이었다. 이 초등학교 학생들의 IQ 분포가 정규분포를 이룬다고 가정할 때, IQ가 100에서 110사이인 학생의 비율은 얼마나 될까?

정규분포의 확률계산 (예제 3) - 문제

- 그렇다면 IQ 가 120 이상인 학생의 비율은 얼마나 될까?

Chapter 3. 분석의 기본 단계 (과정) 알기

- 표본과 표집분포
 - 중심극한정리
 - 통계적 추정
 - 신뢰도와 신뢰구간
 - 통계적 가설검정
 - 유의수준과 P-value
-

기본적인 분석 단계

- 분석하고자 하는 대상에서 표본을 추출 (모집단)
- 표본의 특성 알아냄 (통계량)
- 표본의 특성을 통해서 모집단의 특성을 유추함 (모수)
- 모집단의 특성을 바탕으로 예측 문제를 풀 (각종 분석들)
- 보고서 작성 (끝)

3.1. 표본과 표집분포 (Sample & Sample dist.)

- 표본추출의 필요성
- 오차의 종류
- 표본 (Sample) & 통계량 (Statistic)
- 표집분포

3.1.1. 표본추출의 필요성

- 경제성
- 시간의 제약

3.1.2. 오차의 종류

1. 측정오차 (Measurement Error)
 - 측정하는 과정에서 발생하는 오류 (물리적 오류)
2. 표본추출오차 (Sampling Error)
 - 모집단을 대표할 수 있는 전형적인 구성요소를 표본으로 선택하지 못해서 발생
 - 원인
 - 표본의 크기 때문에 생기는 우연에 의한 오류
 - 편의에 의한 오류

3.1.3. 표본과 통계량 (sample & statistic)

- 통계량 (표본의 특성)

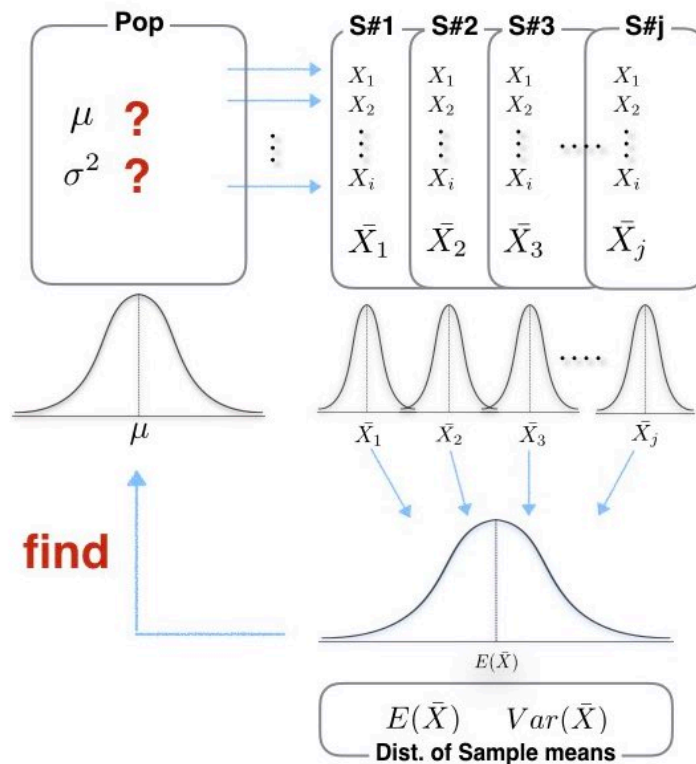
$$\begin{aligned} \text{평균} \quad \bar{X} &= \frac{\sum X_i}{n} \\ \text{분산} \quad S^2 &= \frac{\sum (X_i - \bar{X})^2}{n - 1} \\ \text{표준편차} \quad S &= \sqrt{S^2} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}} \end{aligned}$$

3.1.4. 표집분포 (sampling distribution)

- 표집분포란 모집단에서 일정한 크기로 뽑을 수 있는 표본을 모두 뽑았을 때 그 모든 표본의 특성치, 즉 통계량의 확률분포를 말한다.
 - 모집단에서 표본을 뽑아 그 표본을 분석할 때, 우리가 뽑은 표본이 과연 모집단을 대표할 수 있는가 ?
 - 이는 표본이 포함하고 있는 오차를 추정해 낼 수 있다는 것을 의미 (표집분포가 가능하게 해줌)
 - 똑같은 크기를 가진 표본을 여러 번 추출 -> 각 표본의 특성치인 통계량들 역시 분포를 갖게 됨 -> 이때 통계량이 어떤 분포를 이루는가를 보여주는 것이 표집분포

평균의 표집분포

- 특정한 모집단에서 동일한 크기로 가능한 모든 표본을 뽑아서 각각의 표본들의 평균을 계산했을 때, 그 평균들의 확률분포를 말한다.



모집단이 정규분포일 때

- 평균의 표집분포는 표본의 크기 n 에 관계없이 언제나 정규분포를 이룬다.

$$\text{평균} \quad E(\bar{X}) = \mu$$

$$\text{분산} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

3.1.7. 모집단이 정규분포가 아닐 때

- 모집단이 정규분포가 아닐 때에는 표집분포가 정규분포라고 단정지을 수 없다.
- 그러나 아래의 그림에서와 같이 표집분포는 표본의 크기 n 을 크게 할수록 정규분포에 접근하게 된다. 이를 **중심극한정리 (central limit theorem)**라 한다.

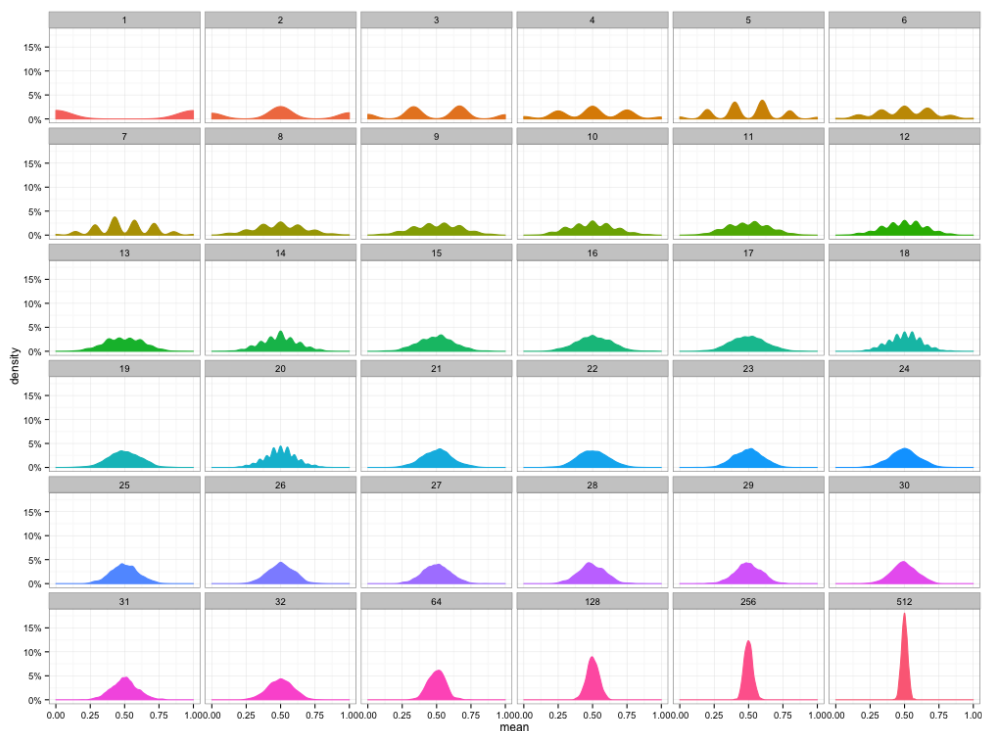
$$\text{평균} \quad E(\bar{X}) = \mu$$

$$\text{분산} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

중심극한정리 (Central Limit Theorem)

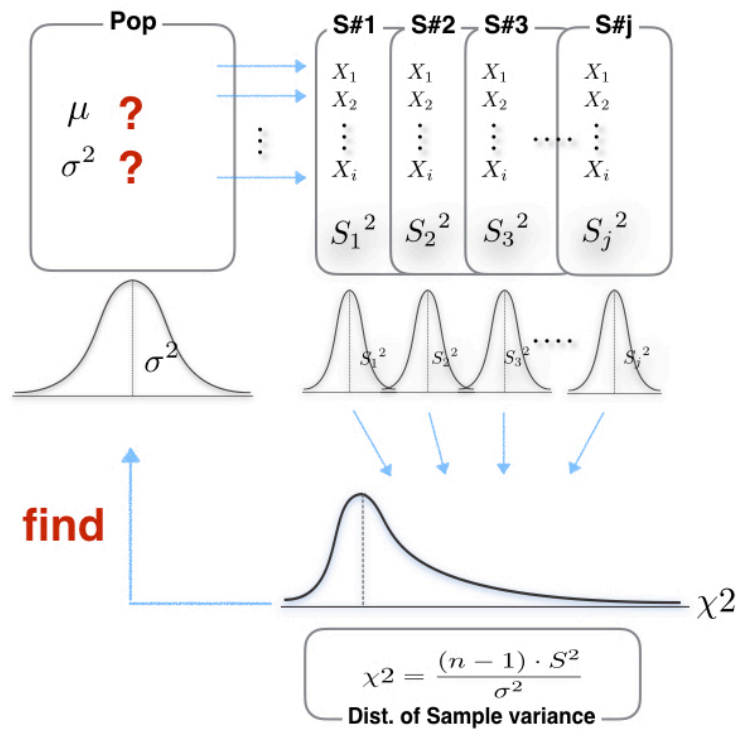
- 통계학에서 가장 중요한 정리
- 모집단의 분포모양과는 상관없이 표본의 크기가 커지면 표집분포가 정규분포를 이루게 되어, 정규분포의 성질을 쉽게 이용할 수 있다는 장점
- 모집단이 정규분포가 아니더라도 n 이 커질수록 정규분포에 접근
 - 대개 표본의 크기가 30 이상이면 정규분포를 이룬다.

중심극한정리 (Central Limit Theorem)



분산의 표집분포

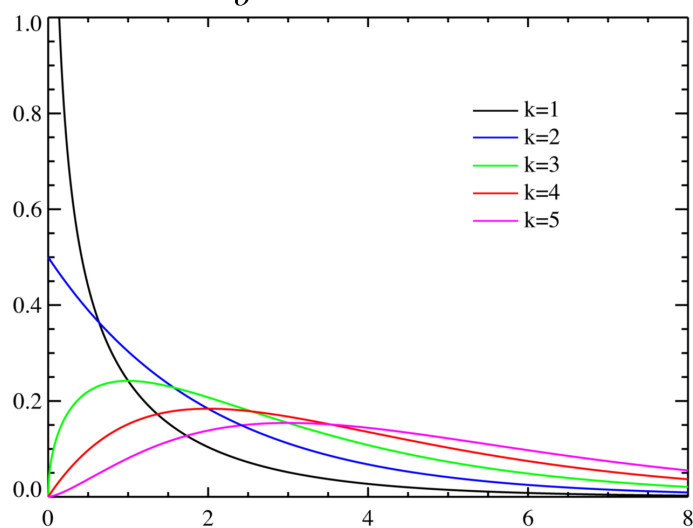
- 어떤 모집단이 σ^2 의 분산을 가질 때, 이 모집단으로부터 크기가 동일하게 선택가능한 모든 표본을 뽑아서 각각의 분산을 계산했을 때,
- 표본분산 S^2 들은 일정한 분포를 이루게 된다.



χ^2 분포(chi-square distribution)

- χ^2 분포는 비대칭 모양을 이루고 오른쪽으로 긴 꼬리를 가짐
- 항상 양수값만을 갖는 특징
- 자유도가 커질수록 (샘플의 크기 n) 정규분포에 가깝게 된다

$$\chi_{n-1}^2 = \frac{(n-1) \cdot S^2}{\sigma^2}, \quad (n-1 = k : \text{자유도})$$



3.2. 통계적 추정

- 통계적 추정의 기본개념
- 점추정
- 구간추정
- 신뢰도와 신뢰구간

Quiz

- 서울시 대학 신입생의 수학능력시험 평균 점수를 알아보려 한다. 각 대학에서 총 400명의 표본을 뽑아 그들의 점수를 조사하여 본 결과 평균 점수가 250점이었다고 한다면, 모집단의 평균 점수는 얼마라고 볼 수 있는가?

A. 250점 일 것이다.

B. 225~275점 일 것이다.

C. 170~330점 일 것이다.

Quiz (답)

- 위 세 개의 답을 보면 어느 하나도 꼭 틀렸다고 말할 수 없다.
 - **A의 경우** 표본의 평균이 250점이라고 해서, 모집단의 평균이 꼭 250점이라고 할 수는 없다.
 - **B의 답**, 즉 225점 이상 275점 이하의 어느 점수가 모수가 될 것이라는 것은 모집단의 평균이 꼭 250점이라는 것보다 맞을 가능성이 높다.
 - **C의 답**, 즉 170점 이상 330점 이하에 모수가 있을 가능성은 거의 100%라고 할 수 있다.

3.2.1. 통계적 추정의 기본개념

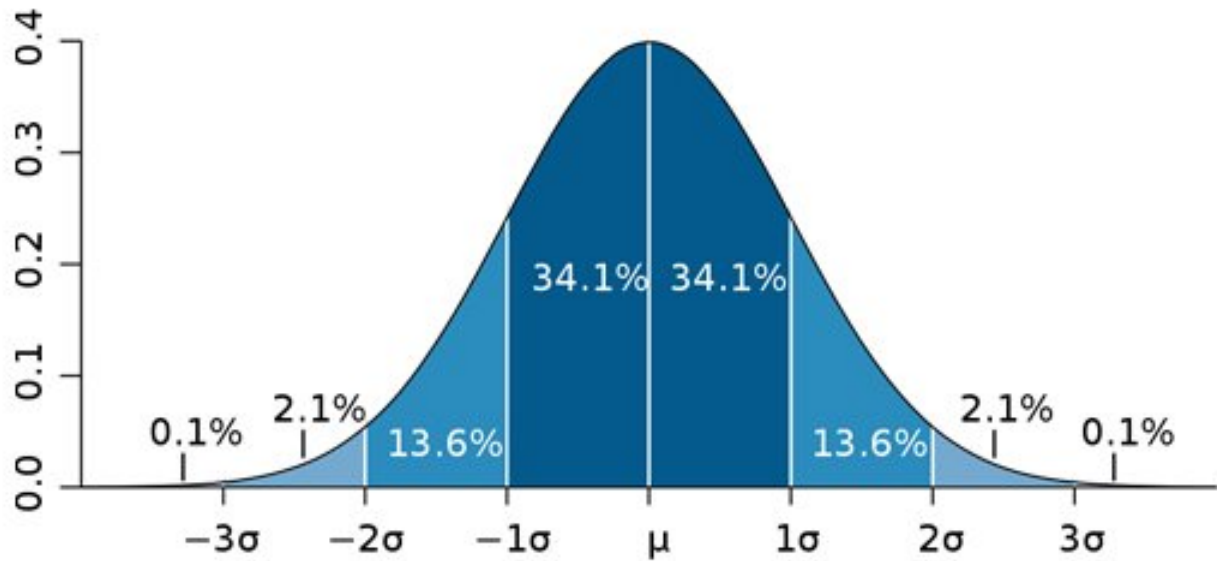
- 정보의 효과와 추정구간의 크기는 상반관계(trade-off)
 - 구간을 크게 할 수록 그 추정이 맞을 가능성은 높아지나, 구간이 클수록 그 정보의 효과는 감소된다.
 - 극단적으로 시험점수가 400점 만점일 때 평균 점수가 0점에서 400점 사이에 있을 것이라고 한다면 그 추정이 맞을 가능성은 100%이지만 그것이 주는 정보의 가치는 하나도 없다.

3.2.2. 점추정 (point estimation)

- 점추정이란 하나의 값으로 모수값을 추정하는 방법이다.
- 다시 말해서, 표본으로부터 구할 수 있는 통계량 가운데 모수를 추정하기에 가장 적절한 것을 결정하여 그 값을 모수값으로 보는 것
 - ex. 평균이 250점 일 것이다.

3.2.3. 구간추정 (interval estimation)

- 구간추정은 모수가 존재할 범위를 제공함으로써 연구자가 원하는 만큼의 정확도를 가지고 모수를 추정할 수 있다는 장점을 가짐



3.2.4. 신뢰도와 신뢰구간

- 신뢰도 또는 신뢰수준 (confidence level)
 - $1 - \alpha$
 - 구간으로 추정된 추정값이 실제 모집단의 모수를 포함하고 있을 **가능성(확률)**
- 신뢰구간 (confidence interval)
 - 이때 모수가 포함될 것으로 추정된 구간
- 신뢰도가 높을수록 신뢰구간은 넓어진다.
- 범위가 넓을수록 그 속에 모집단의 평균이 포함될 가능성이 더 높아지나,
- 반면에 신뢰구간이 갖는 정보의 가치는 줄어들게 됨을 의미

Quiz

- 위의 학생들의 시험점수 문제에서, 표본의 평균 점수가 250점이고, 표준편차가 30점 이라고 할때, 95%의 신뢰구간을 구하여 보자. (위 정규분포 그림을 이용)

3.2.5. 모집단 평균의 구간추정 (σ 를 알고 있는 경우)

Z-통계량

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}}$$

Z 값에 대한 신뢰구간

$$P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1 - \alpha$$

μ 값에 대한 신뢰구간

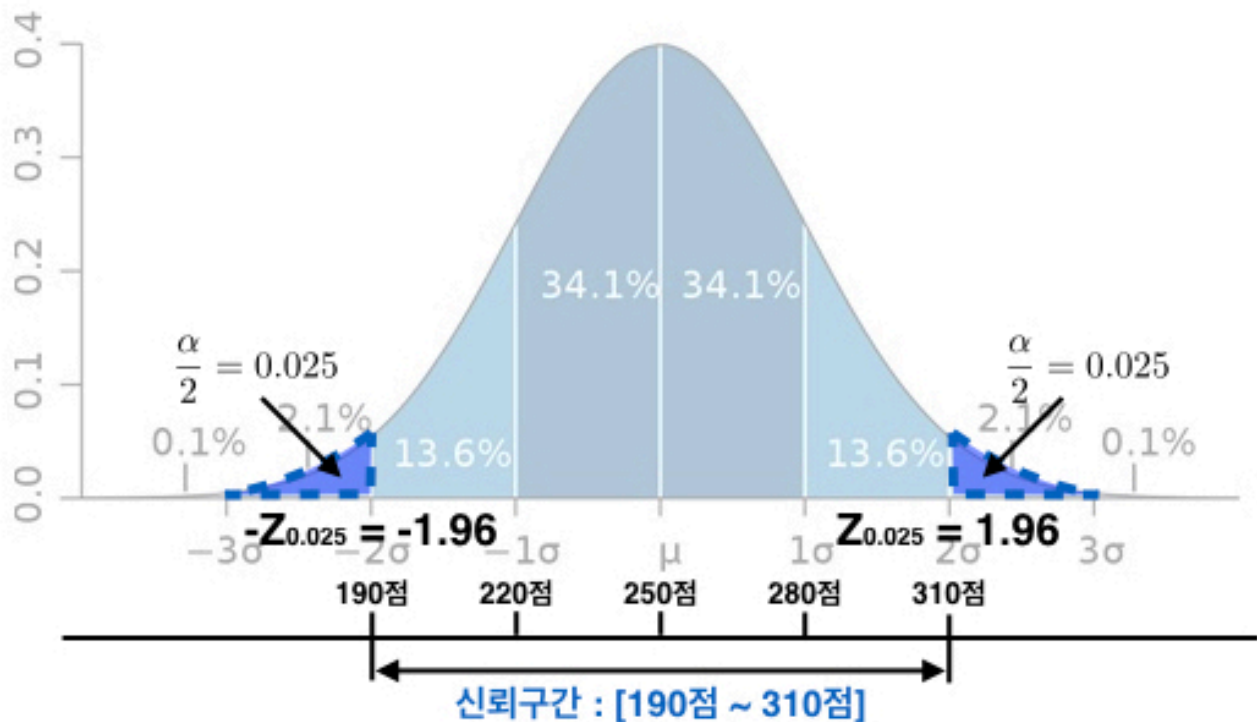
$$P(\bar{X} - Z_{\alpha/2} \cdot \sigma_{\bar{X}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \cdot \sigma_{\bar{X}}) = 1 - \alpha$$

table. 신뢰도에 따른 $Z_{\alpha/2}$ 값

신뢰도 ($1 - \alpha$)	$Z = 0$ 에서 $Z_{\alpha/2}$ 까지 면적	$Z_{\alpha/2}$
0.90	0.450	1.64
0.95	0.475	1.96
0.99	0.495	2.57

Quiz 에 적용

- 위의 학생들의 시험점수 문제에서, 표본의 평균 점수가 250점이고, 표준편차가 30점 이라고 할때, 95%의 신뢰구간을 구 하여 보자.



$$P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1 - \alpha$$

$$P(\bar{X} - Z_{\alpha/2} \cdot \sigma_{\bar{X}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \cdot \sigma_{\bar{X}}) = 1 - \alpha$$

3.2.6. 모집단 평균의 구간 추정 (σ 를 모르는 경우)

- 앞에서 모집단의 평균을 추정할 때 정규분포 모집단의 표준편차를 σ 를 알고 있는 것으로 가정했으나, 모집단의 평균을 모르면서 모집단의 표준편차를 알고 있는 경우는 매우 드물다.
- 대개의 경우 모집단의 표준편차는 모집단의 평균 μ 를 알아야 계산할 수 있기 때문이다.
- 모집단의 표준편차 σ 를 모를 때에는 **표본에서 구한 불편추정량 S , 즉 표본의 표준편차**를 모집단의 표준편차 σ 대신으로 사용한다.

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$$

3.2.6. 모집단 평균의 구간 추정 (σ 를 모르는 경우)

t -통계량

- t -통계량은 표준정규분포를 따르지 않고 자유도 $(n - 1)$ 의 t -분포를 이루기 때문에 t -분포를 이용하여 신뢰구간을 구해야 한다.

$$t = \frac{(\bar{X} - \mu_{\bar{X}})}{S_{\bar{X}}}, \quad (S_{\bar{X}} = \frac{S}{\sqrt{n}})$$

t -통계량을 이용한 신뢰구간

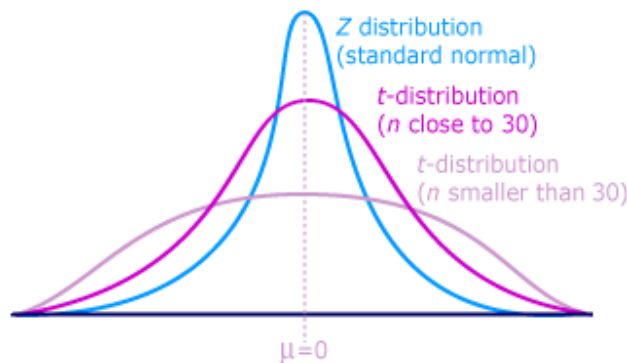
$$P(-t_{\alpha/2} \leq \frac{(\bar{X} - \mu_{\bar{X}})}{S_{\bar{X}}} \leq t_{\alpha/2}) = 1 - \alpha$$

t -분포에서의 신뢰구간

$$P(\bar{X} - t_{\alpha/2} \cdot S_{\bar{X}} \leq \mu \leq \bar{X} + t_{\alpha/2} \cdot S_{\bar{X}}) = 1 - \alpha$$

참고) t -분포에 대해서 (1)

- t -분포의 모양은 Z -분포와 유사
- 종 모양으로서 $t = 0$ 에 대하여 대칭을 이룸
- t -분포는 표준정규분포보다 두터운 꼬리를 가지고 옆으로 퍼져 있음



- 이러한 특성으로 인해 α 에 대하여 t -통계량이 Z -통계량보다 큰 값을 갖도록 함
- 모집단의 표준편차 σ 를 알지 못하는 데서 오는 추정상의 오류를 보상에 준다. (후하게 쳐준다)
- 그러나 표본의 크기 n 이 커질수록 표본의 표준편차 S 는 σ 에 접근하기 때문에 t -분포는 점차 표준정규분포와 비슷한 형태를 이루게 된다. ($n \geq 30$)

데이터 스튜디오를 활용한 기술통계와 데이터 시각화 (실습)

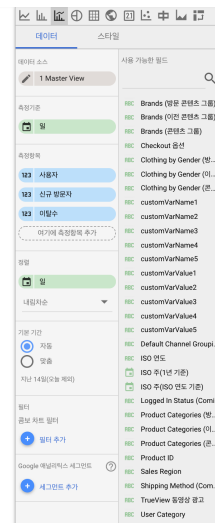


Google Data Studio

- 목표
 - 기본적인 GA 대쉬보드 만들기
 - 조건부확률 실습예제 시각화 하기
- 준비사항
 - GA데모 계정
 - 데이터 스튜디오

데이터 스튜디오의 기본구조

- 데이터 소스
- 측정기준 : dimension
- 측정항목 : metric
- 세그먼트 : segment



수업끝