사회통계
연속확률분포

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 정규분포
2. 표준정규분포
3. 당부사항

정규분포

정규분포

이제부터 연속확률분포에 대해 살펴보자.

- 지금까지 이산확률변수(e.g., 여성별로 출산한 아이의 수)에 관한 이산확률분포를 다루었다.
- 베르누이 분포와 이항분포를 대표적인 이론적 이산확률분포로 학습했다.
- 지금부터 연속확률변수(e.g, 지역별 평균출산율)에 관한 연속확률분포를 다루기로 한다.
- 그 첫번째는 정규분포(normal distribution)가 된다!

정규분포

정규분포는 모든 이론적 확률분포 중에 가장 중요하다.

- Carl Friedrich Gauss (1777-1855)가 발견하여 가우스 분포(Gaussian distribution)라고도 불리운다.
- 탁월한 수학적 아름다움을 갖추었을 뿐만 아니라, 자연과 사회에서 나타나는 셀 수 없이 많은 현상들을 터무니없이 잘 설명한다. “이건 잘 설명하지 못하네” 싶은 현상도 각도를 달리하면 다시 잘 설명한다.
- 무엇보다도 통계적 추론(statistical inference)의 초석 역할을 한다.

독일화폐 10마르크(지금은 유로화로 대체)에는 가우스와 정규분포가 그려져 있다

정규분포

실제 자료를 통해 정규분포를 살펴보자.

- heights.csv 파일을 엑셀에 불러와 살펴보자. 이것은 우리나라 16세 여성 266명을 표본으로 신장을 측정한 것이다.
- 이 자료에서 평균(mean), 중앙값(median), 최빈값(mode)을 엑셀에서 계산하고 비교해보자.
- 키를 확률변수로 생각해보면 이것의 자료유형은 무엇인지 생각해보자.
- 키의 확률분포를 시각화해보자. 계급구간의 수는 대략 20개 정도로 하자. 어떤 특징을 가지고 있는가?

정규분포

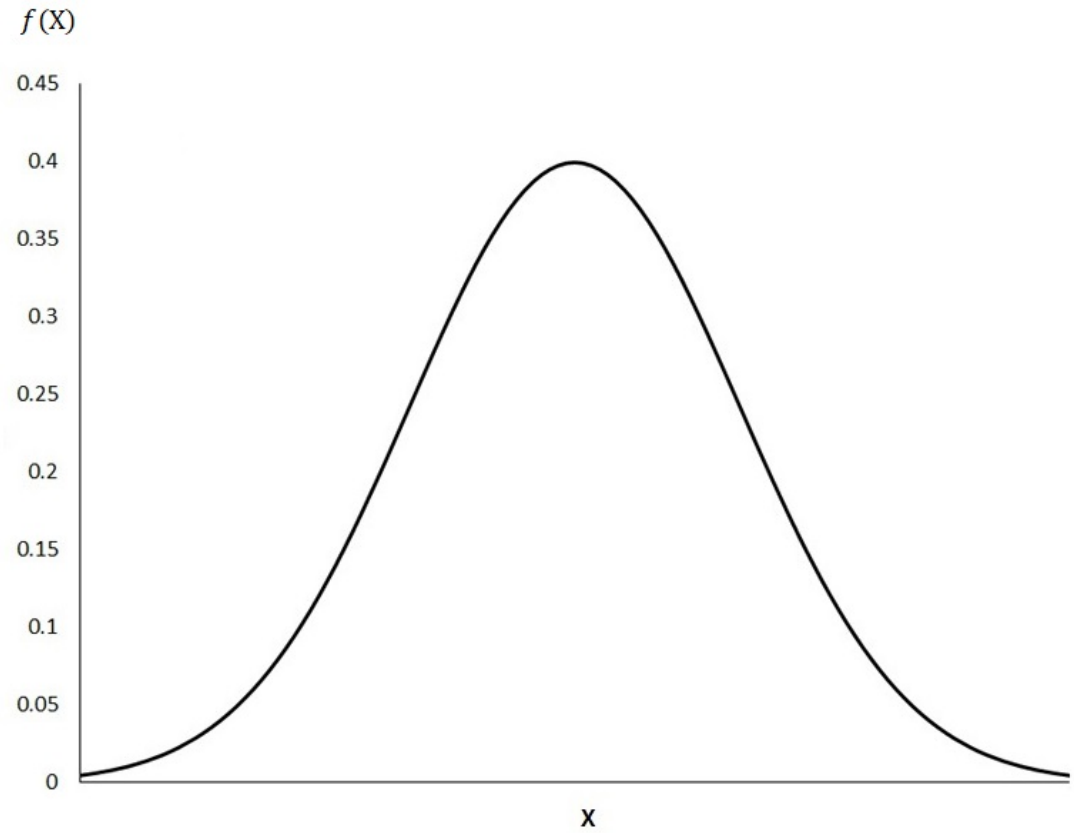
정규분포의 몇가지 기본적인 속성은 다음과 같다.

- 정규분포는 종 모양(bell-shaped)이고 대칭적(symmetric)이다.
- 이론상 양측의 꼬리는 무한히 이어지며 결코 x축에 닿지 않는다.
- 평균, 중앙값, 최빈값은 하나의 정점에서 모두 동일하다.
- x축은 확률변수 X이고 y축은 확률밀도함수(PDF)이다.
- 확률변수 X가 (평균이 μ이고 분산이 σ인) 정규분포를 따른다면 이렇게 표현한다.

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

정규분포

- 정규분포는 아래처럼 그릴 수 있지만, 구체적인 꼴은 (정해진 것이 아니라) μ와 σ에 따라 달라질 수 있음에 주의하자!



정규분포

- 정규분포의 확률밀도함수는 아래와 같다.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- 외우지 말자. 그냥 몇 번이고 종이 위로 옮겨 적어보면서 아름다움을 감상하자.
- 하지만 중요한 감상의 포인트는 있다. 패러미터(parameter)가 무엇인지 기억하는 것이다(μ 와 σ).
- 원주율 π가 나오지만 이것은 상수(약 3.14)이므로 패러미터가 아니다.
- 여기서 e는 이른바 자연상수(natural constant) 또는 오일러 상수(Euler's constant)이다. 이름 그대로 상수(약 2.71)이므로 이것도 패러미터가 아니다.

정규분포

정규분포의 평균과 분산은 무엇일까?

- 일단 확률분포가 주어지면 평균과 분산을 알 수 있다. 정규분포 또한 그렇다.
- 답은 각각 μ와 σ^2이다. 이것은 패러미터로 입력되어야 한다.
- 다만 이것은 약간의 증명이 필요하다. 어렵지는 않지만 이 수업에서는 생략한다.

정규분포

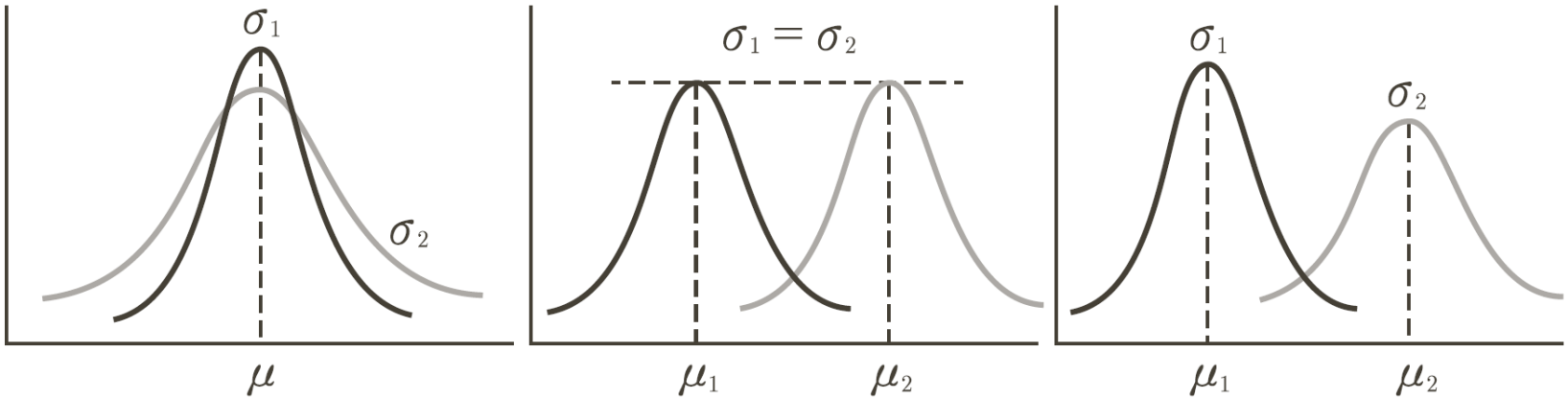
정규분포를 따르는 확률밀도함수를 이해하기 위해 엑셀로 연습하자.

- 엑셀에서 NORM.DIST(x, μ, σ, FALSE) 함수로 정규분포를 따르는 확률밀도함수를 계산할 수 있다.
- normdist.xlsx을 엑셀에 불러오자.
- 어떤 패러미터들이 사용되는지, 또 그것들을 어떻게 확률밀도함수를 계산하는데 어떻게 사용되고 있는지 확인해보자.

정규분포

패러미터에 따라 정규분포의 꼴이 달라진다.

- 평균 μ가 커질수록 그래프의 모양이 어떻게 변하는가?
- 표준편차 σ가 커질수록 그래프의 모양이 어떻게 변하는가?



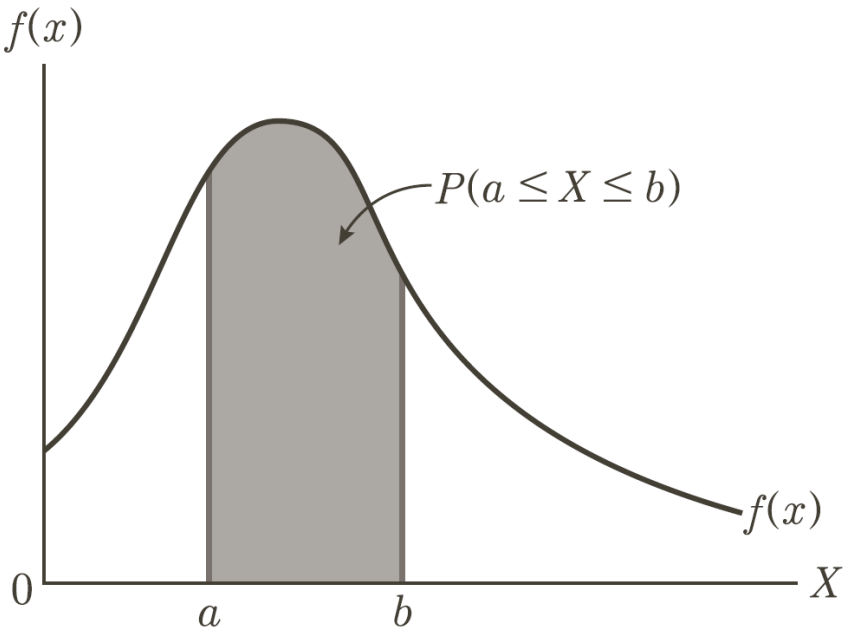
$(\mu_1 = \mu_2, \sigma_1 < \sigma_2)$ $(\mu_1 < \mu_2, \sigma_1 = \sigma_2)$ $(\mu_1 < \mu_2, \sigma_1 < \sigma_2)$

정규분포

정규분포에서도 색칠공부는 존재한다.

- 확률분포를 공부할 때 누적분포함수(CDF)를 함께 다루었듯, 정규분포에서도 누적분포함수를 상정할 수 있다.
- 곡선 아래 면적(area under curve)은 X 범위 안에 놓일 확률을 의미한다(Why?).

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = \int_a^b f(x)dx$$



정규분포

예제 3. 어느 여자중학교 학급에서 신체검사를 수행하여 키의 평균 160cm과 표준편차 5를 얻었다. 여자중학생의 키는 대체로 정규분포한다고 알려져 있다. 키가 156cm 이하인 사람은 전체의 퍼센트인가? 키가 160cm 이상이고 170cm 이하인 사람은 전체의 퍼센트인가?

정규분포

- 이때 확률밀도함수 $f(x)$는 곧 $P(a \leq X \leq b)$임을 기억하자(Why?).
- 그 값은 엑셀 함수 NORM.DIST(x, μ, σ, TRUE)로 구할 수 있다.
- "키가 156cm 이하인 여학생은 전체의 퍼센트인가?"

$$P(X \leq 156) = \int_{-\infty}^{156} f(x)dx = 0.211855399$$

- "키가 160cm 이상이고 170cm 이하인 여학생은 전체의 퍼센트인가?"

$$P(160 \leq X \leq 170) = \int_{160}^{170} f(x)dx = 0.477249868$$

표준정규분포

표준정규분포

여러 정규분포들을 비교하려면 먼저 표준화해야 한다.

- 정규분포의 구체적인 꼴은 패러미터 μ와 σ에 따라 다르다.
- 당연히 서로 다른 여러 정규분포들을 비교하기가 어렵다(왜 비교하는지 곧 공부한다).
- 그러므로 통계학자들 사이에서 $\mu = 0$, $\sigma = 1$로 하는 정규분포 꼴을 이제부터 표준(standard)으로 삼자고 규약을 정했다.
- 그게 바로 표준정규분포(standard normal distribution)다.

표준정규분포

- 왜 하필 $\mu = 0, \sigma = 1$ 일까? 그건 정규분포를 따르는 확률밀도함수를 곰곰히 들여다보면 쉽게 알 수 있다.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- $\mu = 0, \sigma = 1$을 대입하면 자연스럽게 확률밀도함수가 매우 단순해진다(Why?).

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

표준정규분포

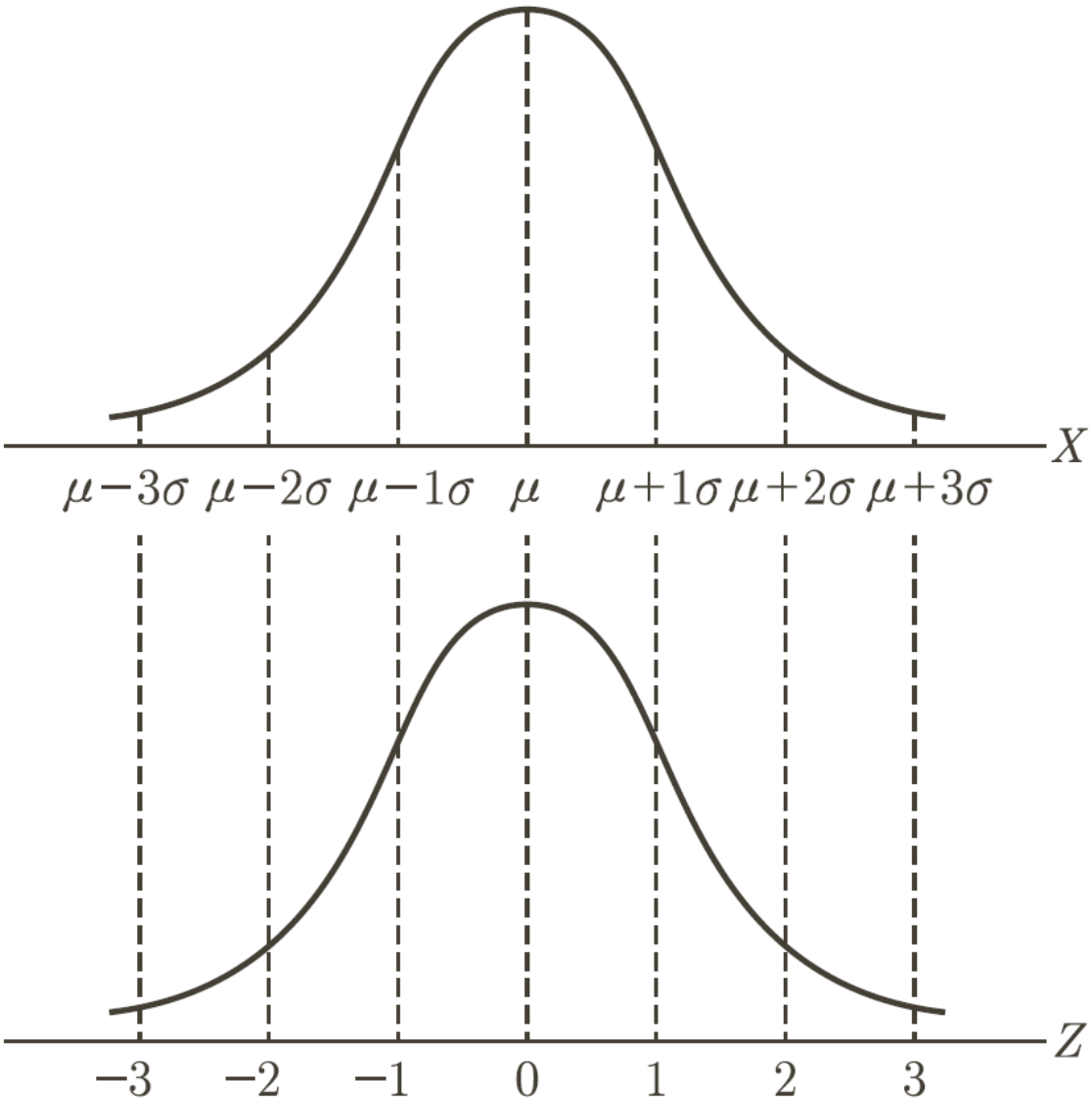
정규분포하는 확률변수를 아주 쉽게 표준화할 수 있다.

- 원점수(raw score)를 Z-점수(Z-score)로 변환하면 어떤 정규분포이던지 표준정규분포로 바꿀 수 있다.
- 원점수 X에서 평균 μ를 빼고 다시 표준편차 σ로 나누어 각각의 Z-점수를 구한다.

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

- 표준화하면 모든 확률변수는 저절로 $\mu = 0$와 $\sigma = 1$이 된다(Why?).

표준정규분포



$$X \sim N(\mu, \sigma^2)$$

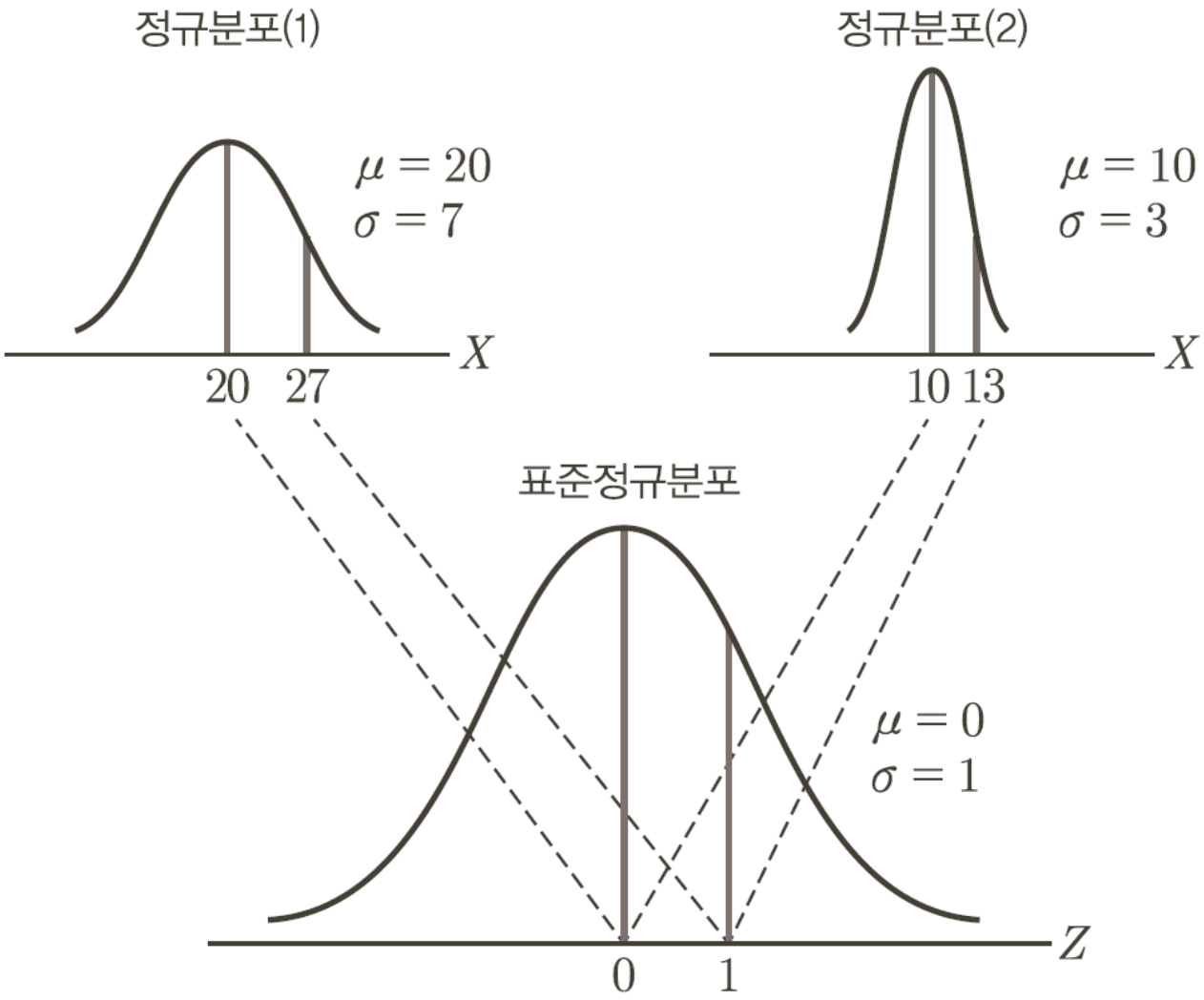
$$Z = \frac{(X-\mu)}{\sigma} \sim N(0,1)$$

표준정규분포

위 그림에서 주의할 점이 있다!

- 표준정규분포는 원래의 정규분표와 보통 그 꼴이 서로 다르다(Why?).
- 꼴을 결정하는 것은 평균 μ와 표준편차 σ였는데 그걸 바꾸었으니 당연하다.
- 위 그림에서는 대응한다는 것을 보여주기 위해 다소 억지로 같게 그린 것임에 주의하자!

표준정규분포


정규분포(1)
$\mu = 20$
$\sigma = 7$
X
20 27
정규분포(2)
$\mu = 10$
$\sigma = 3$
X
10 13
표준정규분포
$\mu = 0$
$\sigma = 1$
Z
0 1

표준정규분포

정규분포하는 확률변수를 실제로 표준화해보자.

- 아까 살펴보았던 heights.csv를 다시 엑셀로 불러오자.
- 평균 μ와 표준편차 σ를 한 켠에 계산해두고 표준화 공식에 따라 각각의 원점수를 표준화하자(셀 참조를 할 때 평균 셀과 표준편차 셀은 고정해야 한다).

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

- (원점수 X와 마찬가지로) 이렇게 계산한 표준화된 값 Z 역시 연속확률변수이다. 그러므로 이것의 확률분포도 다시 표와 그래프(히스토그램)로 나타낼 수 있다.

표준정규분포

예제 4. 어느 병원의 의무기록지에 따르면 이 병원에서 지난 5년 간 태어난 신생아의 몸무게 평균은 3.2kg였고 표준편차는 0.7kg이였다. 이 병원 자료에 따르면 신생아의 출생 몸무게는 정규분포한다. 이 병원에서 어젯밤 새벽 새우라는 이름의 한 아이가 태어났다.새우의 출생시 몸무게는 4.1kg 이였다. 이 병원에서 태어난 아이 중 새우보다 가벼운 아이는 전체의 몇 퍼센트인가?

표준정규분포

- 먼저 원점수를 사용하여 곧바로 계산해보자.
- NORM.DIST(4.1, 3.2, 0.7, TRUE) 엑셀 함수를 사용하면 약 0.901을 얻는다.
- “전체 출생아의 90.1%는 새우보다 출생 몸무게가 가볍다.”
- 이번엔 Z-점수로 표준화하여 계산해보자.
- NORM.DIST(Z, 0, 1, TRUE) 엑셀 함수를 사용해도 똑같은 값을 얻는다(Why?).
- 이때 Z 값은 다음과 같이 계산한다(Why?).

$$Z = \frac{X - \mu}{\sigma} = \frac{4.1 - 3.2}{0.7} = 1.285714$$

표준정규분포

표준정규분포하는 자료를 어떻게 본래의 정규분포로 되돌릴 수 있을까?

- 간단하다! Z-점수를 원점수로 환원하면 된다.
- Z-점수에서 표준편차 σ를 곱한 값에 평균 μ를 더하여 원점수를 구한다(Why?).

$$X = Z \cdot \sigma + \mu$$

- 이건 앞서 소개한 표준화 공식을 다시 X에 대해 정리한 것에 불과하다.

표준정규분포

예제 5. 작년 정부 보고서에 따르면 공직자적성검사의 점수는 평균 72점과 표준편차 8점으로 정규분포한다고 알려졌다. 하위 40% 이하는 무조건 탈락이라고 한다. 과락을 면하려면 최소한 몇 점을 넘어야 하나?

표준정규분포

- 이 문제는 (처음에 했던 연습과는 반대로) 확률이 주어졌을 때 그에 대응하는 원점수를 묻고 있다.

$$P(X < ?) = 0.4$$

- 대충 찍어보자! 한 60점 나오면 하위 40% 이하가 되어 과락 아닐까?
- 엑셀 함수로 NORM.DIST(60, 72, 8, TRUE)를 입력해보면 약 0.067을 얻는다. 하위 6.7%이므로 너무 낮다!
- 그러면 한 70점 나오면 어떨까?
- 엑셀 함수로 NORM.DIST(70, 72, 8, TRUE)를 입력해보면 약 0.401을 얻는다. 하위 40.1%이므로 정답에 매우 근접했다.
- 그러나 매번 찍기를 반복할 수는 없다.

표준정규분포

- 다행히 확률을 주면 그에 대응하는 Z-점수를 찾아주는 NORM.INV(p, μ, σ) 엑셀 함수가 있다.
- 이 함수는 특히 정규분포에서 누적분포함수의 역(inverse)을 계산하는데 쓰인다.
- 가령 $P(Z <?) = .8$ (즉, 상위 20%)를 찾고 싶다면, NORM.INV(.8, 0, 1)를 입력한다.
- 즉, NORM.DIST(·)와 NORM.INV(·)는 서로 역관계인 셈이다.
- NORM.DIST(·)에서는 Z-점수를 넣고 확률을 얻었다면, NORM.INV(·)에서는 확률을 넣고 Z-점수를 얻는다.
- 다만 NORM.DIST(·)와는 달리, 어차피 누적분포함수를 전제로 하므로 FALSE나 TRUE는 넣을 필요없다(무조건 TRUE다).

표준정규분포

- 아까 문제로 되돌아가 찍지말고 제대로 다시 풀어보자.
- 엑셀 함수 NORM.INV(.4, 0, 1)를 사용하면 약 -0.25라는 Z-점수를 얻는다.
- 그런데 이 Z-점수만 봐서는 뭐가 뭔지 알 수 없고, 요걸 다시 원점수로 환산해야 의미를 파악할 수 있다(Why?).
- 이 실수가 매우 흔하기 때문에 주의해야 한다!
- 아까 배운 식을 활용해서 $-.25 \cdot \sigma + \mu = 70$을 얻을 수 있다.
- 정말로 약 70점을 넘어야 하위 40% 이상인 셈이다.

당부사항

당부사항

당분간 (시간 제한은 똑같은데) 사회통계 수업이 제법 어려워진다.

- 수업 시간에 연습한 엑셀 파일은 저장하여 집에 가져가자. 단기 기억이 쇠퇴하기 전에 빨리 복습하여 자신만을 위한 기록을 남기자.
- 강의안 뿐 아니라 반드시 교재를 구비하여 관련 챕터를 공부해야 한다. 몇 주 전 내용이라도 복습해야 한다.
- 이 수업에서는 대표적인 이산확률분포와 연속확률분포만 공부했지만 교재를 통해 좀 더 공부해야 한다.
- 특히 이산균등분포(discrete uniform distribution)와 포아송 분포(Poisson distribution)를 반드시 공부해야 한다.

6주차 과제

문제 1번(이항분포만), 2번, 3번, 4번 (150페이지)
(4번 문제에서 여성의 표본 수는 300명임!)