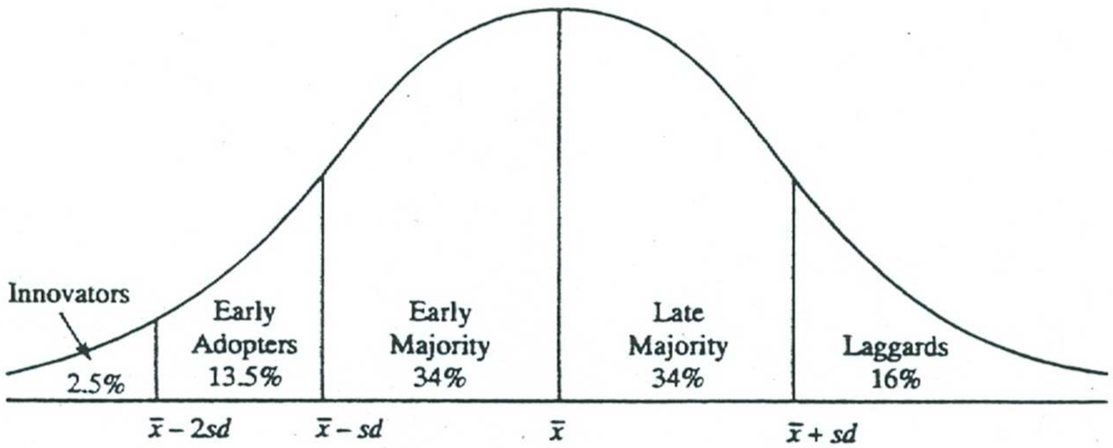
통계분석 기법의 선택

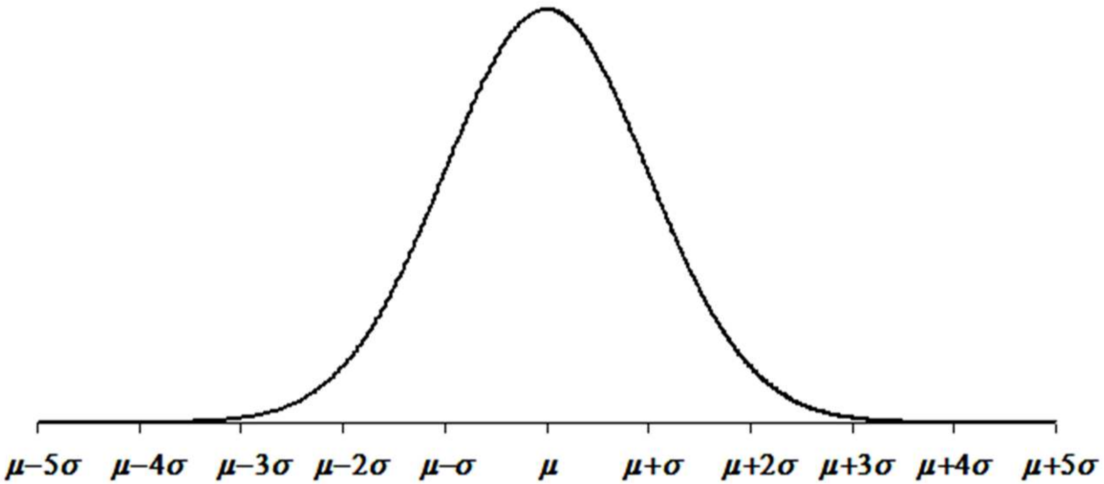
사례 : 혁신적인 제품의 확산과 정규분포

신제품이 출시되었을 때, 소비자들마다 그 제품을 채택하는 시기가 다르다. 어느 소비자들은 시장에 나오자마자 구매하는 반면, 어떤 소비자들은 한 참이 지난 후에야 구매한다. 왜 이러한 차이가 존재하는 것일까? 이러한 문제를 처음 체계적으로 연구한 것은 사회학자 에버렛 로저스(Everett M. Rogers)였다. 그는 농부들이 옥수수 씨를 새로운 품종의 개종 씨앗으로 바꾸는 확산과정에 관한 연구를 수행하였다. 이 연구결과를 토대로, 소비자들의 특성에 따라 혁신을 수용하는 시기가 다르다는 사실을 하나의 이론으로 체계화하였다. 이것이 기술수용주기(technology adoption life cycle, TALC) 모형이다.

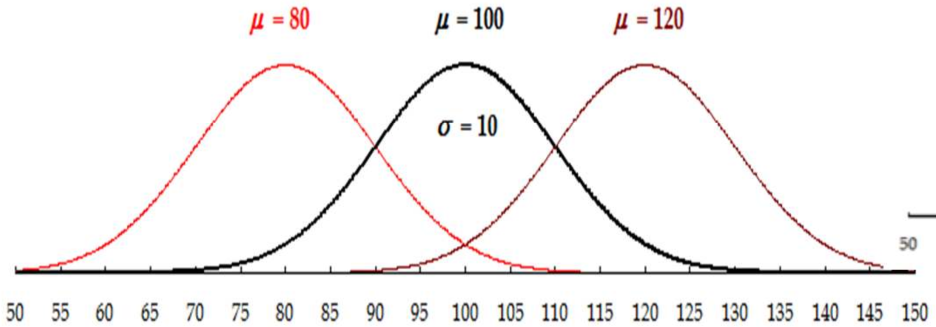


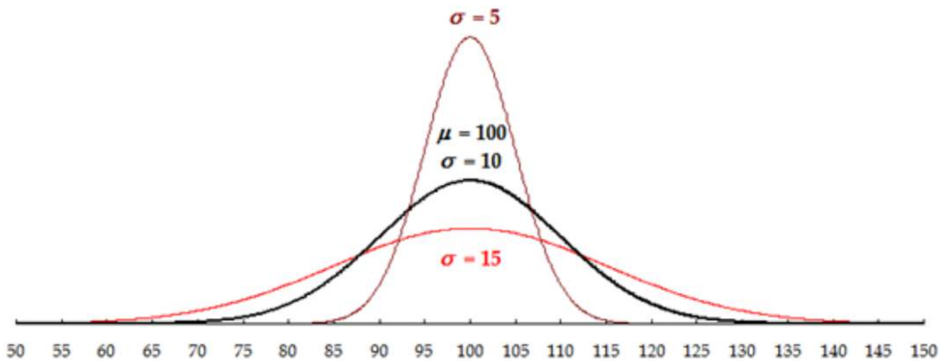
정규분포의 특징과 확률계산

◆ 정규분포 곡선



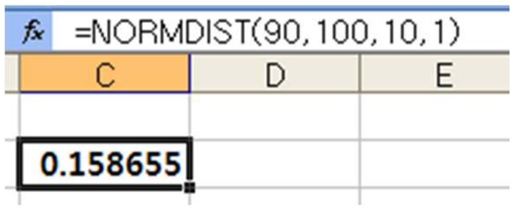
◆ 평균과 표준편차의 변화에 따른 정규분포곡선의 변화





정규분포의 특징과 확률계산

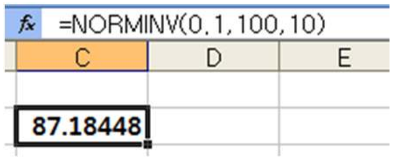
술을 사랑하는 50명의 애주가들의 모임이 있다. 그들에게 있어 현실이란 알코올 결핍 때 나타나는 환영에 불과하다. 아일랜드의 탐미주의자 오스카 와일드(Oscar Wilde)가 말한 것처럼, 애주가들에게 직업이란 저주이다. 하지만 정말 바쁠 때는 술을 안 먹으니 알코올 중독자는 아니다. 이들이 한 번 모임에서 마시는 소주 양이 정규분포를 따른다고 하자. 평균은 소주 100병이고 표준편차는 10병이다.
이들이 모임에서 소주를 90병보다 적게 마실 확률을 구해보자. 즉 $Pr (X < 90)$을 구하기 위해서는, 엑셀의 셀에 다음과 같이 입력하면 된다.

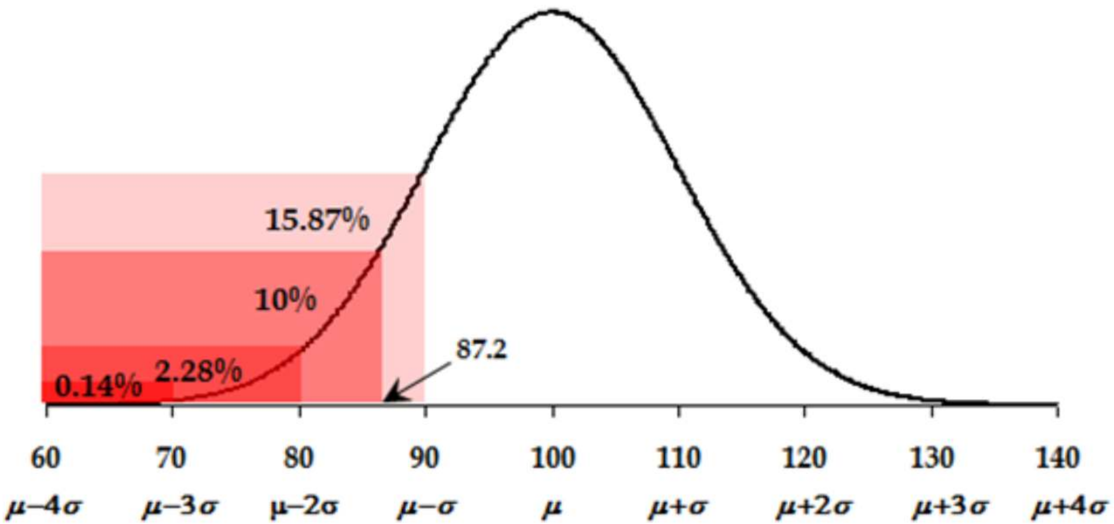


평균보다 적게 마시는 수준		확률값	백분율
90병보다 적게 마실 확률	$Pr (X < 90)$	0.1586552539	15.87%
80병보다 적게 마실 확률	$Pr (X < 80)$	0.0227501319	2.28%
70병보다 적게 마실 확률	$Pr (X < 70)$	0.0013598980	0.14%
60병보다 적게 마실 확률	$Pr (X < 60)$	0.0000316712	0.00%
50병보다 적게 마실 확률	$Pr (X < 50)$	0.0000002867	0.00%
40병보다 적게 마실 확률	$Pr (X < 40)$	0.0000000098	0.00%

정규분포의 특징과 확률계산

$Pr (X < x) = 0.1$인 x를 구하기





$Pr (X < x) = 0.2$인 x를 구하기???

표준화와 표준정규분포

술의 종류는 정말 많다. 소주, 막걸리, 맥주, 양주, 칵테일, 폭탄주 등등… 요즈음은 막걸리가 대 유행이다. 하루는 동료가 “맙(막)소사주”를 마시자고 해서 그게 무엇이냐고 물었더니, ‘막걸리+소주+사이다’를 적당한 비율로 섞는 것이라고 한다. 술 마시는 방법도 여러 가지이다. 가끔 우리는 상대방에게 주량을 물어보곤 한다. 하지만 서로 다른 술을 기준으로 사용하기 때문에 비교하기가 어렵다. 소주병의 수로 환산하여 말하는 것이 술꾼들 사이의 예의이다. 예를 들면, 맥주 4병을 소주 1병으로 인정하는 것이다. 이를 단위변환이라고 말한다.

- 표준화 공식
  - 모집단의 경우: $Z = \frac{x-\mu}{\sigma}$
  - 표본의 경우: $Z=\frac{x-\bar{X}}{s}$

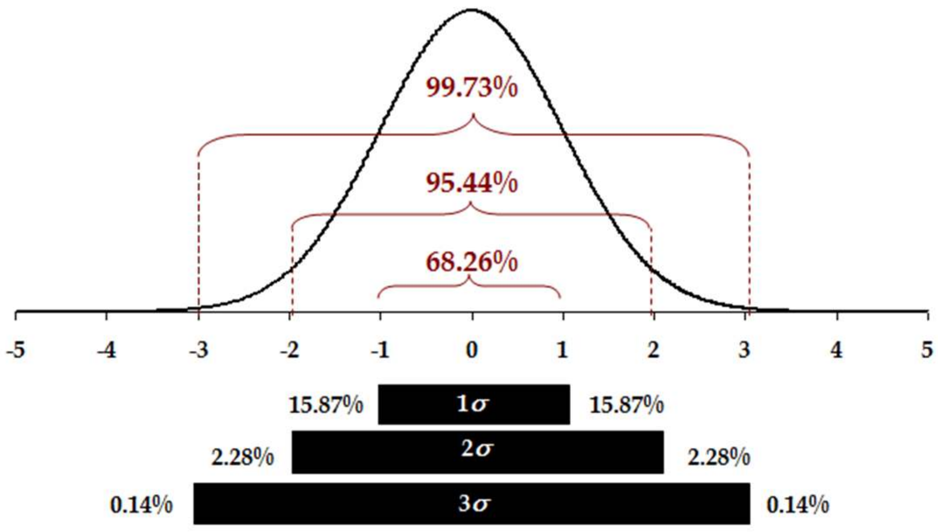
표준화와 표준정규분포

Q: 애주가인 주희(酒喜)씨는 주량이 소주 2병쯤 되어, 그 정도면 기분 좋게 취한다. 맥주로는 한 8병쯤 된다. 그러면 소주 3병을 마시는 것과 맥주 12병을 마시는 것 중 어디 쪽이 더 취하겠는가? 그 날 상태에 따라 좌우되는 주량의 표준편차가 중요하다. 소주 주량의 표준편차는 1병이고, 맥주의 표준편차는 6병이라고 하자.

A:?

표준편차와 확률

◆ **표준편차에 따른 확률**



◆ **한국인 20대 남성 키~N(174, 6^2)**

- 1σ 밖의 값을 관찰할 확률은 32% 정도: 180㎝ 이상이거나 168㎝ 이하인 사람은 32%
- 2σ 밖의 값을 관찰할 확률은 5%도 안됨: 186㎝ 이상이거나 162㎝ 이하인 사람은 5% 미만
- 3σ 밖의 값을 관찰할 확률은 0.27% 정도: 192㎝ 이상이거나 156㎝ 이하인 사람은 0.3%도 채 안됨

표준정규분포 확률 구하기

◆ 표준정규분포 확률 구하기 공식

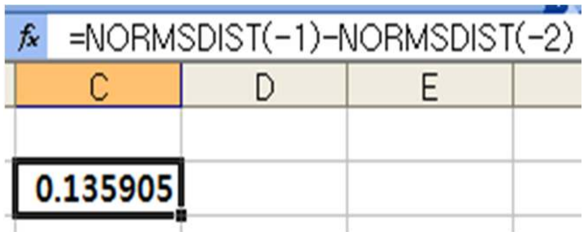
● 확률변수 Z가 평균이 0이고 표준편차가 1인 표준정규분포를 따른다고 하자.

❖ $Pr(Z < z)$를 구할 때, =NORMSDIST (z)

❖ $Pr(Z < z) = p$인 z를 구할 때, =NORMSINV (p)

◆ '조기 수용 집단(early adopter group)'의 비중을 구하기

● $Pr(-2 < Z < -1) = Pr(Z < -1) - Pr(Z < -2)$



죄수의 딜레마: 독립과 아닌 무엇

구소련의 스탈린 시대에, 한 오케스트라 지휘자가 기차를 타고 연주회장으로 향하고 있었다. 그날 밤 지휘할 곡의 악보를 검토하고 있었는데, 이를 지켜보던 두 명의 KGB 비밀경찰은 그를 간첩혐의로 체포하려 했다. 그는 악보가 간첩들이 사용하는 비밀암호가 아니라, 그저 차이코프스키의 바이올린 협주곡일 뿐이라고 항의하였다. 그러나 그의 주장은 묵살되었고, 체르젠스키 광장의 한 KGB 건물로 붙잡혀 가게 되었다. 감금된 지 이틀째 되는 날, 비밀경찰들이 그에게 와서 말하였다. "모든 걸 숨김없이 말하는 것이 좋을 거야. 당신 친구, 차이코프스키도 우리들에게 체포되었어. 그리고 이미 다 털어 놓았으니까."

죄수의 딜레마: 선택대안과 형벌

		차이코프스키: 결백 주장	차이코프스키: 거짓 자백
지휘자	결백 주장	3, 3	25, 1
지휘자	거짓 자백	1, 25	20, 20

<표 4-4> 죄수의 딜레마: 지휘자의 선택

		차이코프스키	
		결백 주장	거짓 자백
지휘자	결백 주장	3 3	1 25
	거짓 자백	25 1	20 20

<표 4-5> 죄수의 딜레마: 차이코프스키의 선택

		차이코프스키	
		결백 주장	거짓 자백
지휘자	결백 주장	3 3	1 25
	거짓 자백	25 1	20 20

죄수의 딜레마: 내쉬의 균형점

		차이코프스키	
		결백 주장	거짓 자백
지휘자	결백 주장	3 3	1 25
	거짓 자백	25 1	20

변수의 측정수준과 통계적 분석

- **가정: 변수 X와 변수 Y의 관련성을 알고자 함**
  - Case 1. 범주형과 범주형
    - 흡연여부와 폐암유병여부의 관련성
    - 변수 *X*는 흡연자인지 비흡연자인지를 나타내고, 변수 *Y*는 폐암환자인지 아닌지를 나타냄
    - 흡연자집단과 비흡연자집단의 발생비율을 비교하여, 이 발생비율의 차이가 충분히 크다면 두 변수는 서로 관련 있음
    - ‘카이제곱 독립성 검정’이나 ‘두 모비율 차이에 대한 검정’을 이용
  - Case 2. 수치형과 수치형
    - 변수 X는 나이를 나타내고, 변수 Y는 근력
    - 한 변수가 증가할 때 다른 변수는 어떻게 변하는지를 살펴, 두 변수 사이의 관련성에 대해 설명
    - 상관관계분석을 이용
  - Case 3. 범주형과 수치형
    - 변수 *X*는 남자인지 여자인지 성별구분을 나타내고, 변수 *Y*는 근력을 나타냄
    - 남성집단과 여성집단의 근력 평균을 비교해서 그 차이가 크다면, 성별구분과 근력은 관련이 있다고 말할 수 있음
    - *T* 검정, 분산분석을 이용

모수적 기법 VS. 비모수적 기법

- ◆ **모수적 기법 : 모집단에 대한 이론적 분포를 가정하고, 모수를 이용해 통계적인 분석을 하는 기법**
- ◆ **비모수적 기법: 모집단에 대한 이론적 분포를 가정하지 않는통계분석 기법**

변수의 측정수준	모수적 기법	비모수적 기법
범주형 – 범주형		카이제곱 독립성 검정 (9장)
(이분형 – 이분형)	두 모비율 차이의 z 검정 (9장)	카이제곱 독립성 검정 (9장)
수치형 – 수치형	피어슨 상관분석 (10장)	
박쥐형 – 박쥐형		스피어만 순위상관계수 (10장)
이분형 – 수치형	t 검정 (11장)	
이분형 – 박쥐형		맨-휘트니 U 검정 (11장)
범주형 – 수치형	분산분석 (12장)	
범주형 – 박쥐형		크러스컬-월리스 검정 (12장)