

Chapter13-1.

통계분석(기술통계 & 확률분포)



목차

1. 통계기법
2. 기술통계
3. 확률분포
4. 정규분포(연속확률분포)
5. 이항분포(이산확률분포)



1. 통계기법

- 통계기법(Statistical Methods)

- ✓ 자료에서 유용한 정보를 추출하고, 패턴을 식별하여 의사결정 및 예측 수행

- 통계기법 분류

1. **기술통계**(Descriptive Statistics): 자료를 요약하고 설명하는 데 사용된다. 통계 지표는 평균, 중앙값, 분산, 표준편차, 최솟값, 최댓값, 사분위수 등
예) A회사 전직원의 연령에 대한 중심 경향성과 분포 파악
2. **추론통계**(Inferential Statistics): 표본 자료를 이용하여 모집단의 모수(평균, 분산 등)를 추정하는 데 사용된다. 신뢰구간 추정, 가설검정, 회귀분석 등 포함
예) 모집단의 평균에 대한 가설검정으로 표본 평균이 모집단 평균과 차이 확인



추론통계 영역

● 추론통계(Inferential Statistics) 영역

- ✓ 표본으로 모집단의 특성 또는 모집단의 확률분포에 대한 결론을 도출 하기 위해서 사용되는 통계학의 한 분야
- 1. 신뢰구간 추정 : 모집단의 모수가 포함될 신뢰구간을 추정
- 2. 가설검정 : 모수로 가설을 정하고, 가설의 사실 여부를 통계적으로 검증
 - t 검정(t-test) : 모평균과의 차이 검정
예) 성인 여성의 평균키와 표본 키와 차이가 있는지 검정
 - 분산분석(Analysis of Variance): 그룹 간의 통계적 차이를 분석하는 데 사용
예) 제품 성능을 결정하는 처리방식을 그룹 간에 비교 및 처리방식 판단
 - 카이제곱검정(Chi-Square Test): 범주형 자료 간의 독립성을 평가하고, 관련성 분석에 사용한다. 예) 부모의 학력수준과 자녀의 대학진학 여부 관계
 - 회귀분석(Regression Analysis): 독립변수와 종속변수 간의 관계를 모델링하고 예측하는 데 사용 예) 광고비와 매출액 간의 인과 관계를 분석하고 예측



2. 기술통계

- 기술통계 관련 모듈 : **import statistics**

Basic statistics module.(기본 통계 함수)

Function	Description
mean	Arithmetic mean (average) of data.
fmean	Fast, floating point arithmetic mean.
geometric_mean	Geometric mean of data.
harmonic_mean	Harmonic mean of data.
median	Median (middle value) of data.
median_low	Low median of data.
median_high	High median of data.
median_grouped	Median, or 50th percentile, of grouped data.
mode	Mode (most common value) of data.
multimode	List of modes (most common values of data).
quantiles	Divide data into intervals with equal probability.

Calculating variability or spread(산포도 관련 함수)

Function	Description
pvariance	Population variance of data.
variance	Sample variance of data.
pstdev	Population standard deviation of data.
stdev	Sample standard deviation of data.



2. 기술통계

- 기술통계 관련 모듈 : `import scipy.stats`

stats : Statistical Functions (통계함수)

왜도 : 오른쪽 or 왼쪽 치우침 척도
`scipy.stats.skew(x)`

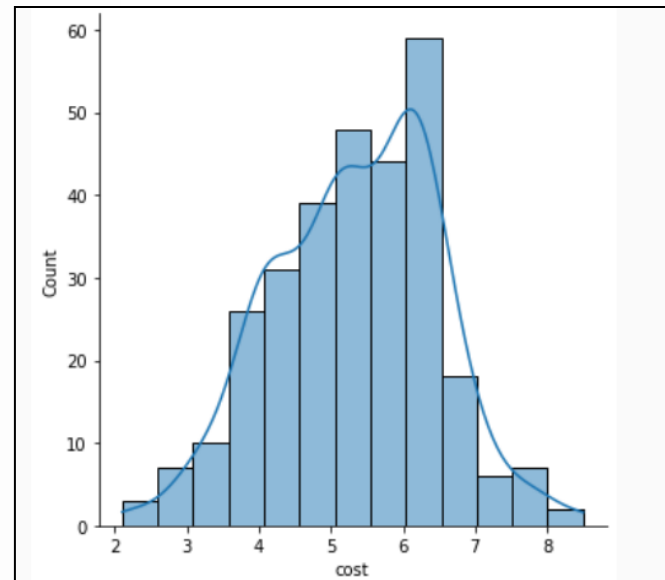
첨도 : 가장 높은 봉우리 척도
`scipy.stats.kurtosis(x)`

히스토그램 + 밀도분포곡선
`import seaborn as sn`

`sn.displot(x, kind='hist', kde=True)`
[해설] 밀도분포곡선으로 왜도와 첨도 확인

정규분포 : 왜도 = 0

정규분포 : 첨도 = 3(Pearson기준)





2. 기술통계

1) 중심 측도 : 평균

- 중심에 대한 측도로 가장 많이 사용하는 것이 평균이다. n 개의 데이터가 x_1, x_2, \dots, x_n 일 때 평균

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum x_i$$

- 이상점이 있을 때 평균은 믿을 만 한가?

다음은 2014년 프로야구선수 14명의 연봉(단위: 천만원)자료이다.

↓ 15.078

2.4 2.4 2.6 2.7 2.8 2.8 2.8 3.6 4.0 10.0 20.0 25.0 30.0 100.0

평균을 계산하면 15.078 즉 15,078만원이다. 가장 많은 연봉 10억원을 제외하고 계산하면 8,546만원이다. 연봉 하위 8명의 평균은 2,800만원이다. 이와 같이 **이상점**들이 존재할 때 평균은 믿을 만한 중심 측도가 아닐 수 있다.



3



2. 기술통계

3) 변동성의 측도 : 평균을 중심으로 얼마나 퍼져 있는지를 나타내는 지표

- **사분위수(quartile)** : 데이터를 크기 순으로 정렬했을 때 4등분하는 숫자
사분위수 3개 중에서 가장 작은 사분위수를 제1사분위수라고 하고 Q1
으로 표기하고 Q1보다 작은 데이터는 전체의 25%이다.
가장 큰 사분위수는 제3사분위수라고 하고 Q3로 표기하며, Q3보다 큰
데이터는 전체의 75%이다. 제2사분위수는 중앙값(median)이다.

예) 2014년 프로야구팀의 14명의 연봉(단위: 천만원)이 다음과 같을 때 사분위수

2.4 2.4 2.6 2.7 2.8 2.8 2.8 3.6 4.0 10.0 20.0 25.0 30.0 100.0

↓ ↓ ↓ ↓

mean 15.0785

Q1 25% 2.675

Q2 50% 3.200

Q3 75% 21.250



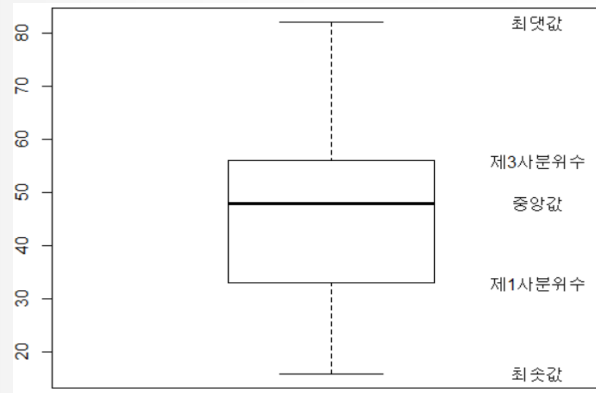
2. 기술통계

● 사분위수와 상자그래프

➤ 데이터 분포를 다섯 숫자로 요약하면 다음과 같다.

1. 최솟값
2. Q1
3. 중앙값
4. Q3
5. 최댓값

상자 그래프(box plot)

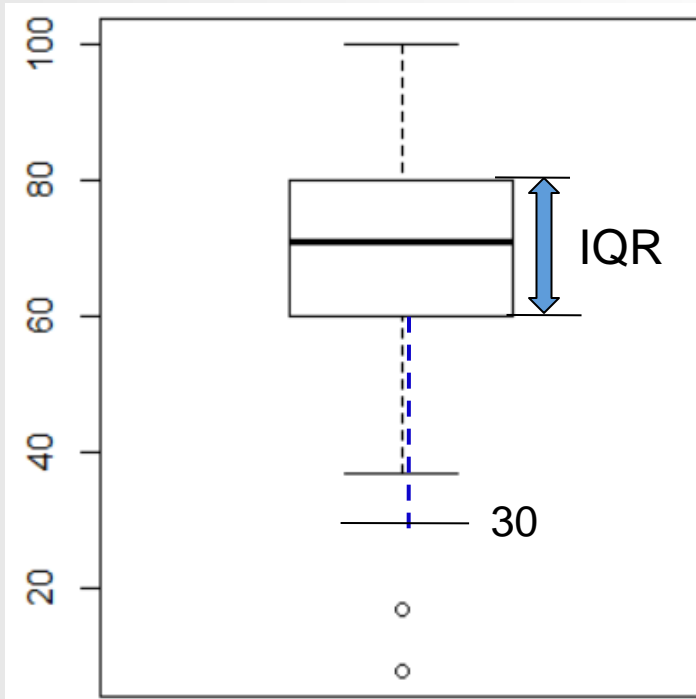


상자 양쪽에 최솟값과 최댓값으로 수염(whisker)을 그린다.



2. 기술통계

● 상자그래프와 이상점



48명 중간고사 시험 점수

8 17 37 44 46 48 51 53 55 57 57 60 60 60 62 64 64 64 64 64
66 68 68 71 71 71 71 73 73 75 75 77 77 77 77 80 80 80 84 84
84 86 88 88 95 95 95 100

사분위수 범위(IQR) = $Q3 - Q1$
수염의 최대 길이는 사분위수 범위의 1.5배까지 가능하다.

수염의 최대길이를 벗어나지 않는 범위에서 가장 작은 값과 가장 큰 값을 찾아 수염을 연결한다.

제1사분위수 - $1.5 * IQR = 30$

제3사분위수 + $1.5 * IQR = 110$

최대값은 110을 넘지 않으므로 최대값 100까지 수염을 연결한다.

하지만 8과 17은 최솟값 30을 벗어난 값으로 이상점으로 간주한다.

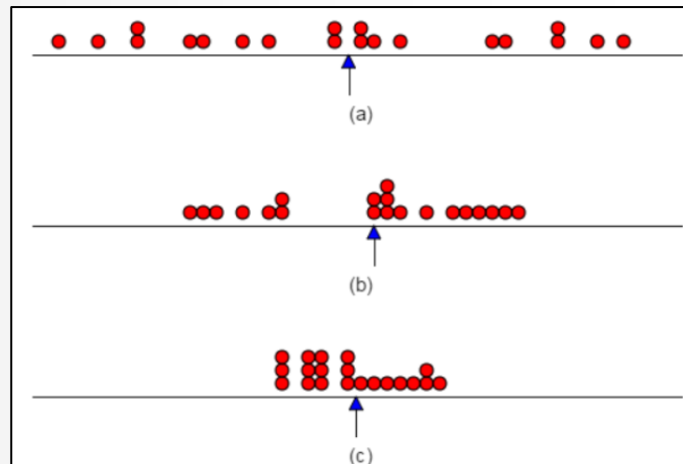


2. 기술통계

3) 변동성의 측도 : 평균을 중심으로 얼마나 퍼져 있는지를 나타내는 지표

- **표준편차** : 평균을 중심으로 자료가 몰려 있는지 아니면 넓게 퍼져 있는지를 측정할 때 분산(variance)과 표준편차(standard deviation)를 사용된다.

데이터가 '한 곳에 몰려 있다' 또는 '넓게 퍼져 있다'라는 표현은 상대적이다. 아래 그림에서 (b)는 (a)보다 몰려 있지만 (c)보다는 넓게 퍼져 있다. 따라서 데이터 집합의 퍼진 정도를 숫자로 표현하는 것이 객관적이다.





2. 기술통계

편차(deviation) : 각각의 자료와 평균과의 거리를 종합하여 하나의 수치로 표현할 수 있다. 각각의 데이터에서 평균을 뺀 값을 편차라고 한다.

분산(variance) : 각 편차를 제곱하여 합한 후에 (데이터 개수-1)로 나눈 값을 **분산(variance)**이라고 하며 S^2 로 표기한다.

표준편차(standard deviation) : 분산에 제곱근을 적용하여 하나의 수치로 표현하며, S 로 표기한다.

평균, 분산, 표준편차를 계산하는 공식

데이터 x_1, x_2, \dots, x_n 이 있을 때, 평균, 분산, 표준편차는 다음과 같이 계산합니다.

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} \\ s^2 &= \frac{\sum (x_i - \bar{x})^2}{n - 1} \\ s &= \sqrt{s^2}\end{aligned}$$



2. 기술통계

● 표준편차의 특징

- ✓ 표준편차는 항상 0보다 크거나 같다. 표준편차가 0일 때는 모든 데이터가 같다는 의미이다.
- ✓ 표준편차의 측정 단위는 원 자료와 같다. 예를 들면 원 자료의 측정 단위가 센티미터이면 표준편차의 측정 단위도 센티미터이다. 그러나 분산의 측정 단위는 제곱 센티미터가 된다. 따라서 변동성의 측도로 분산 보다는 표준편차를 더 많이 사용한다.
- ✓ 표준편차도 평균처럼 이상점의 영향을 크게 받는다. 이상점 몇 개로 표준편차 값이 매우 커질 수 있다.



2. 기술통계

3) 변동성의 측도 : 평균을 중심으로 얼마나 퍼져 있는지를 나타내는 지표

- **변동계수**(Coefficient of Variation) : 표준편차를 평균으로 나눈 값으로 평균이 다른 두 집단 비교 시 유용하다. $CV = \frac{s}{\bar{x}}$

예) 두 회사의 월급 분포

회사	평균 월급 (만원)	표준편차 (만원)
A	300	30
B	200	40

$$A\text{회사} = 30/300 = 0.1$$

$$B\text{회사} = 40/200 = 0.2$$

해석

- B회사의 급여는 평균에 비해 더 많이 변동한다.
- 즉 급여의 일관성은 A회사가 더 높다.
- 값이 작을수록 안정적이고, 클수록 불안정하다.



3. 확률분포

● 확률변수와 확률분포

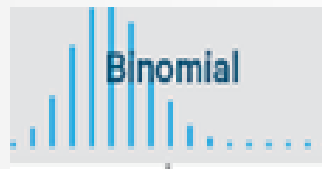
1. 확률변수 : 확률실험에서 사용되는 변수

- ✓ 이산확률변수 : 동전실험, 주사위실험, 불량품개수(셀 수 있음)
- ✓ 연속확률변수 : 시험점수, 키, 몸무게, 노트북사용시간(셀 수 없음)

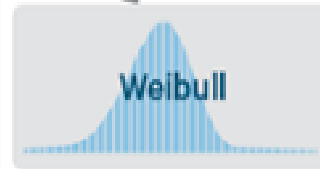
2. 확률분포 : 어떤 확률변수의 확률 분포를 나타내는 함수

- ✓ 이산확률분포 : 확률질량함수(PMF)를 이용하여 변수의 크기와 모양 제공
- ✓ 연속확률분포 : 확률밀도함수(PDF)를 이용하여 변수의 크기와 모양 제공

확률질량함수



확률밀도함수

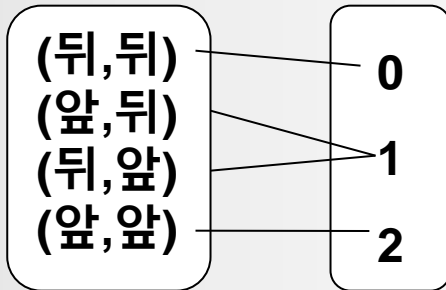




확률질량함수(PMF : Probability Mass Function)

- 동전 2회 던져서 앞면(1)이 나오는 개수를 확률변수 x 라고 할 때 (0:모두실패)

독립시행2회 확률변수 x (질량)



확률질량함수 $P(X=x)$

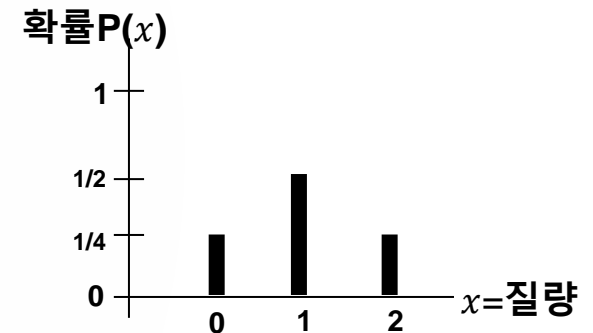
$$P(X=0) = \frac{1}{4} = 0.25$$

$$P(X=1) = \frac{2}{4} = 0.5$$

$$P(X=2) = \frac{1}{4} = 0.25$$

각 질량의 확률 계산

확률변수 x 의 확률분포



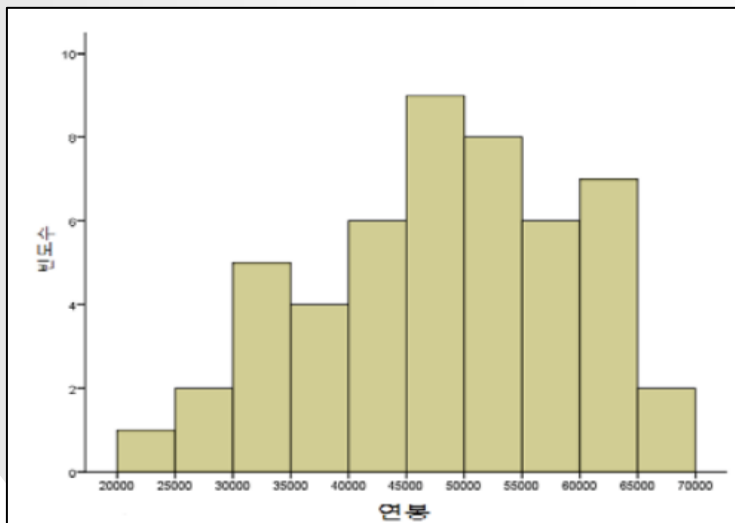
확률변수 x 확률표현 : $P(0 \leq x \leq 2) = P(0) + P(1) + P(2) = 0.25 + 0.5 + 0.25 = 1$

질량함수 성질 : 확률의 전체합 = 1

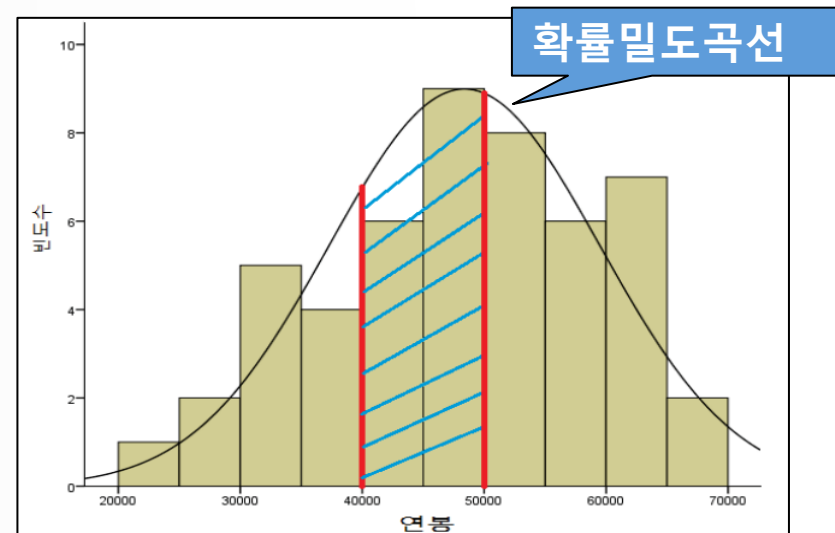


확률밀도함수(PDF : Probability Density Function)

- 어느 연구소의 연구원 50명을 대상으로 한 연봉 자료를 확률변수 x 라고 할 때
 - ✓ 확률밀도곡선 : 계급의 크기를 무한히 작게, 분포다각형으로 그릴 때 그려지는 곡선(곡선의 면적 = 1)
 - ✓ 자료 분포의 대략적인 패턴이나 모양 파악 및 특정 구간 통계적 추론



PDF
➡

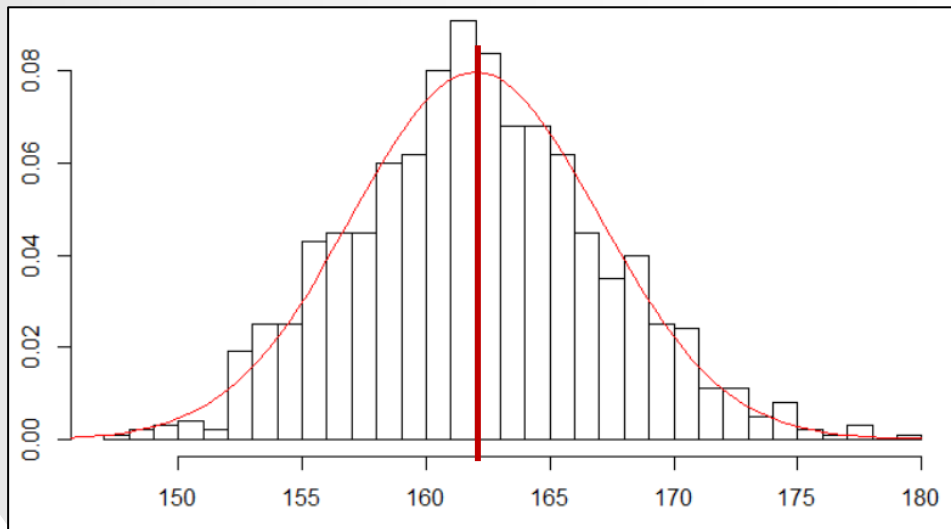




밀도 곡선

● 밀도곡선(Density Curve)

통계청이 제공하는 자료에 따르면 우리나라 성인 여자(19~24)의 평균 키는 162cm, 표준편차는 5cm라고 한다. 다음은 성인 여자 1,000명의 키를 조사한 후에 히스토그램으로 그린 것이다.



밀도 곡선의 두 가지 성질

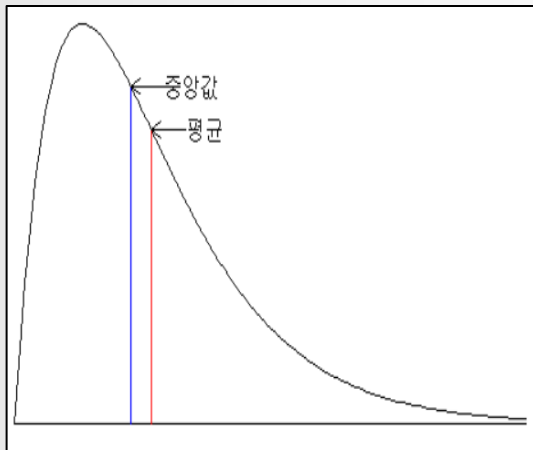
1. 항상 비율 0 이상인 값을 가진다.
2. 밀도 곡선 아래 면적은 항상 1



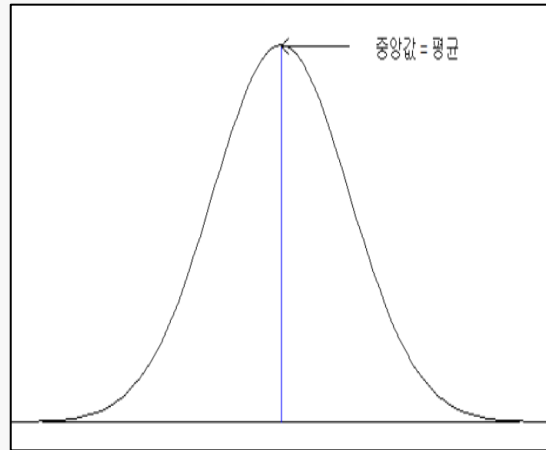
밀도 곡선

● 밀도곡선(Density Curve) 유형

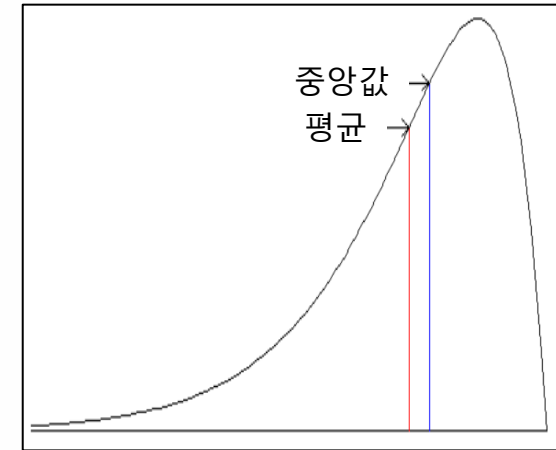
밀도 곡선이 오른쪽으로 길게 늘어진 꼬리를 갖게 되면 중앙값보다 평균이 더 큰 값을 가지게 된다. 길게 늘어진 오른쪽 꼬리에 있는 값들이 평균을 오른쪽으로 끌어 당기기 때문이다.



오른쪽으로 길게 늘어진 꼬리



좌우대칭



왼쪽으로 길게 늘어진 꼬리



확률분포 관련 모듈

- 확률분포와 검정 관련 모듈 : `from scipy import stats`

=====

연속확률분포(Continuous distributions)

`beta.rvs(a=2, b=5, size=100)` : 알파(a)와 베타(b) 구간의 연속확률분포

`chi2.rvs(df=자유도, size=100)` : 극단값을 허용하는 멍분포, 표본 수가 많을수록 대칭

`f.rvs(dfn=분자자유도, dfd=분모자유도, size=100)` `stats.f` : 두 `chi2`분포를 각각의 자유도(d.f)로 나눈 비율을 나타낸 분포

`norm.rvs(loc=모평균, scale=표준편차, size=100)` : 정규분포, 좌우 대칭분포

`t.rvs(df=자유도, size=100)` : 표본수가 작은 경우(30개 미만) 정규분포 대신 사용

`uniform.rvs(a=0, b=1, size=100)` : 0~1사이에서 균등하게 표본추출된 확률분포

=====

이산확률분포(Discrete distributions)

`bernoulli.rvs(p, size=100)` : 독립시행 1번(베루누이시행)으로 추출된 베루누이 분포

`binom.rvs(n=시행횟수, p, size=100)` 독립시행 n번으로 표본이 추출된 이항분포

`geom.rvs(p=성공확률, size=100)` : 최초 성공할 때 까지 실패한 횟수를 갖는 기하분포

`poisson.rvs(mu= 평균발생횟수, size=100)` : 발생 가능성이 매우 작은 포아송분포

=====



연속확률분포

● 연산확률분포 유형

유형	의미
정규분포 (가우스분포)	평균에 가까울수록 발생할 확률이 높고, 멀어질수록 확률이 적은 분포 모양으로 <u>평균값을 중심으로 좌우대칭인 종</u> 모양을 갖는 확률분포
표준정규분포 (Z분포)	정규분포를 대상으로 평균=0, 표준편차=1로 표준화하는 표준화 공식을 이용하여 표준화 시킨 정규분포(모평균 검정)
T분포	표본수가 작고(30개 미만), 좌우대칭 형태를 가지며, 표본수가 많아 질 수록 정규분포와 가까워 지는 확률분포, 정규분포 대신 사용(모평균 검 정)
카이제곱(χ^2)분포	극단값을 허용하는 분포로 표본 수가 많을수록 좌우대칭 분포를 갖다. 표준정규분포에 제곱을 취해서 얻어진 확률분포(모분산 검정)
F분포	두 집단의 카이제곱(χ^2) 분포를 각각의 자유도로 나눈 비율을 나타낸 확률분포(모분산 검정)



이산확률분포

이산확률분포 유형

문제) 다음 보기 중 성격이 다른 확률분포는?

가. 기하분포 나. 정규분포 다. F분포 라. 지수분포

유형	의미
베르누이 분포	두 가지 범주(성공 or 실패)를 갖는 이항분포에서 <u>독립시행 $n=1$인 경우 베르누이 시행</u> 이라 하며, 이 시행으로 얻은 확률분포(동전을 1번 던져서 앞면이 나올 확률(p)은 50%)
이항분포	두 가지 범주(성공 or 실패)를 갖는 이항분포에서 <u>독립시행 n번</u> 으로 각 시행에서 성공확률 p 를 갖는 확률분포(동전을 10번 던져서 앞면이 나올 확률(p)은 45%)
기하분포	베르누이 시행에서 처음 성공하기 전 <u>실패한 횟수</u> 를 갖는 확률분포
음이항분포	베르누이 시행에서 처음 실패하기 전 <u>성공한 횟수</u> 를 갖는 확률분포
포아송분포	시간, 면적 등 지정된 단위 구간에서 특정한 사건의 <u>발생 가능성</u> 이 매우 작은 확률분포(예 : 번개에 맞을 확률)



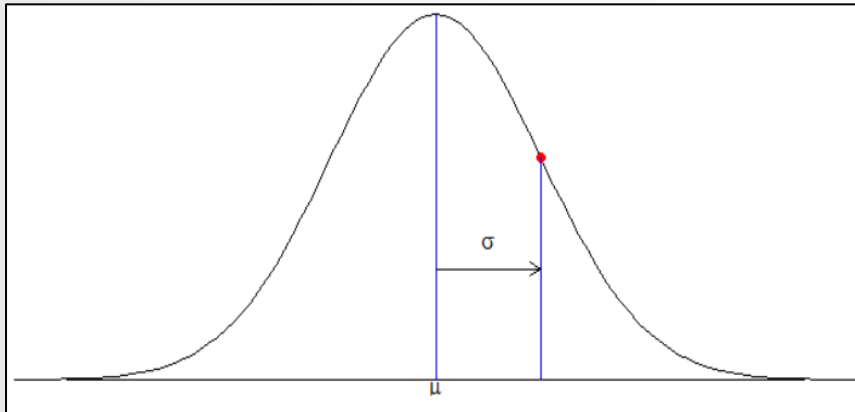
4. 정규분포

● 정규분포

정규분포(normal distribution)는 다음 그림과 같이 좌우 대칭 이고, 봉우리는 한 개이며, 전체적인 모양은 종과 비슷하다.

정규분포는 평균 μ 와 표준편차 σ 에 의하여 결정된다.

표준편차는 정규곡선의 기울기가 변하는 변곡점에 위치하므로 눈으로도 대략 표준편차의 위치를 찾을 수 있다.



키, 체중, 수능 점수 등과 같이 대부분의 데이터가 근사적인 정규분포를 따른다. 하지만 현실의 어떤 데이터도 정확한 정규 분포를 따르지 않는다. 그러나 **근사적인 정규분포를 따른다는 사실만으로 통계에서 발생하는 중요한 문제들을 많이 해결할 수 있다.**



정규성 검정

● 정규성 검정

주어진 데이터가 정규분포를 따르는지 여부를 판단하는 통계적 검정

[사례] A대학의 학생 30명의 시험 성적을 수집하여 이 성적이 정규분포를 따른다고 가정할 수 있는지 알고자 한다.

- 귀무가설 (H_0): 시험 성적은 정규분포를 따른다.(차이가 없다)
- 대립가설 (H_1): 시험 성적은 정규분포를 따르지 않는다. (차이가 있다)

```
from scipy.stats import shapiro
import numpy as np

# 샘플 데이터 (시험 성적 예시)
scores = np.array([85, 90, 88, 93, 95, 87, 91, 89, 84, 86,
                  90, 92, 85, 87, 89, 90, 91, 88, 86, 85,
                  89, 90, 93, 92, 91, 94, 95, 90, 88, 89])

# Shapiro-Wilk 정규성 검정
stat, p = shapiro(scores)
print(f"p-value: {p:.4f}")
```

결과 : p-value: 0.6123
귀무가설 채택

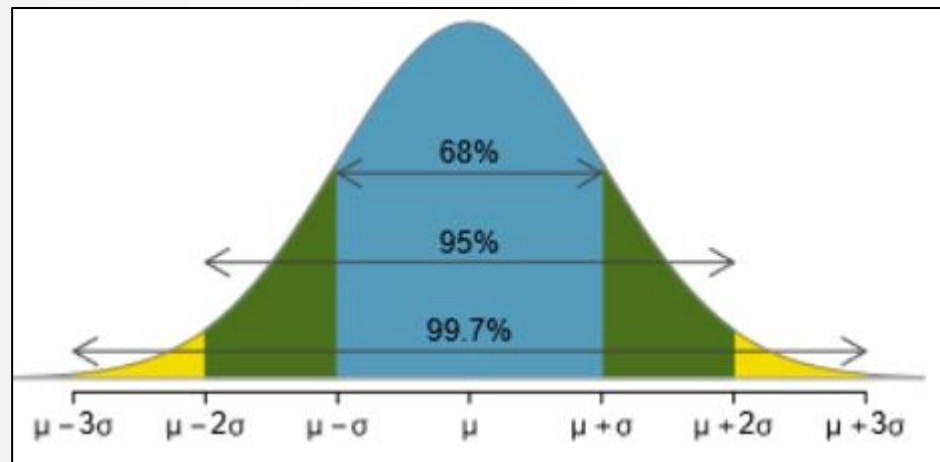


68-95-99.7 법칙

● 68-95-99.7 법칙

정규분포는 평균 μ 와 표준편차 σ 가 변함에 따라 모양이 다르지만 모든 정규분포는 다음과 같은 공통적인 특징을 갖는다.

- ✓전체 데이터 중에 대략 68%는 구간 $(\mu - \sigma, \mu + \sigma)$ 에 있다.
- ✓전체 데이터 중에 대략 95%는 구간 $(\mu - 2\sigma, \mu + 2\sigma)$ 에 있다.
- ✓전체 데이터 중에 대략 99.7%는 구간 $(\mu - 3\sigma, \mu + 3\sigma)$ 에 있다.

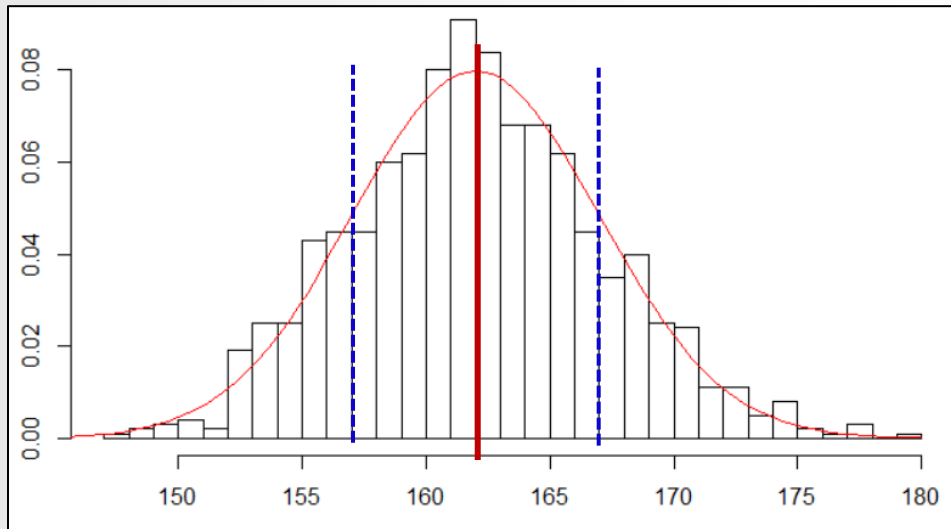




근사적 정규분포

● 근사적인 정규분포의 밀도 곡선

통계청이 제공하는 자료에 따르면 우리나라 성인 여자(19~24)의 평균 키는 162cm, 표준편차는 5cm라고 한다. 다음은 성인 여자 1,000명의 키를 조사한 후에 히스토그램으로 그린 것이다.



빨간색의 밀도 곡선에서 파랑색 구간은 157~167cm 사이로 비율은 68%(680명)와 근사하다.

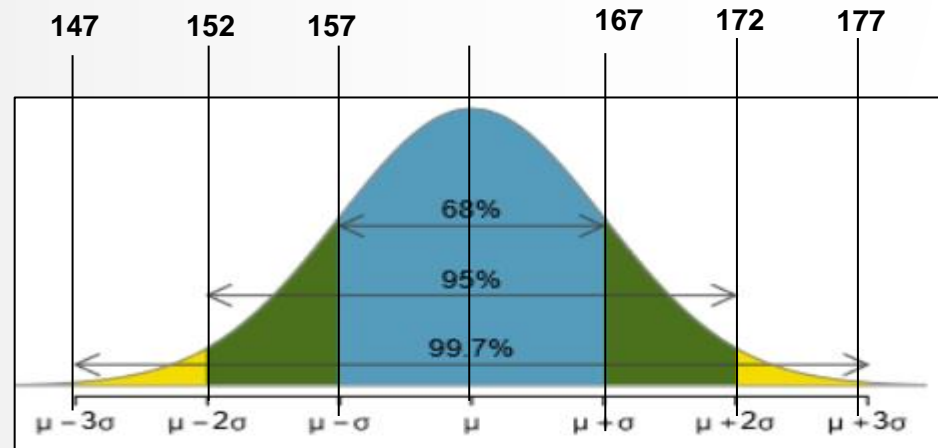
근사적으로 정규분포가 가정될 때 t-검정, 회귀분석 등 다양한 통계 분석에 적용할 수 있다.



근사적 정규분포 사례

[사례] 성인여자(19~24세)의 키는 평균 162cm, 표준편차는 5cm 정규분포를 따른다.

- 1) 키가 157cm에서 167cm 사이에 있는 여자들의 비율은 **68%**
- 2) 키가 167cm 이상인 여자들의 비율은 **16%** (50% - 24%)
- 3) 키가 172cm 이상인 여자들의 비율은 **2.5%** (50% - 47.5%)
- 4) 키가 147cm 이하이거나 177cm 이상인 여자들의 비율은 **3%** (100%-99.7%)이며 1,00명의 여자 중에 대략 **3명** 있다.

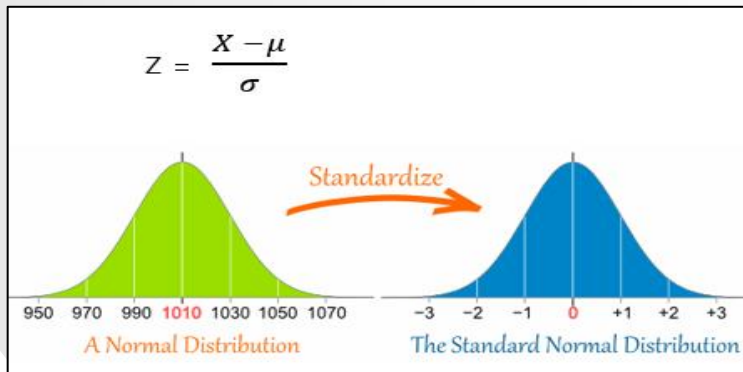




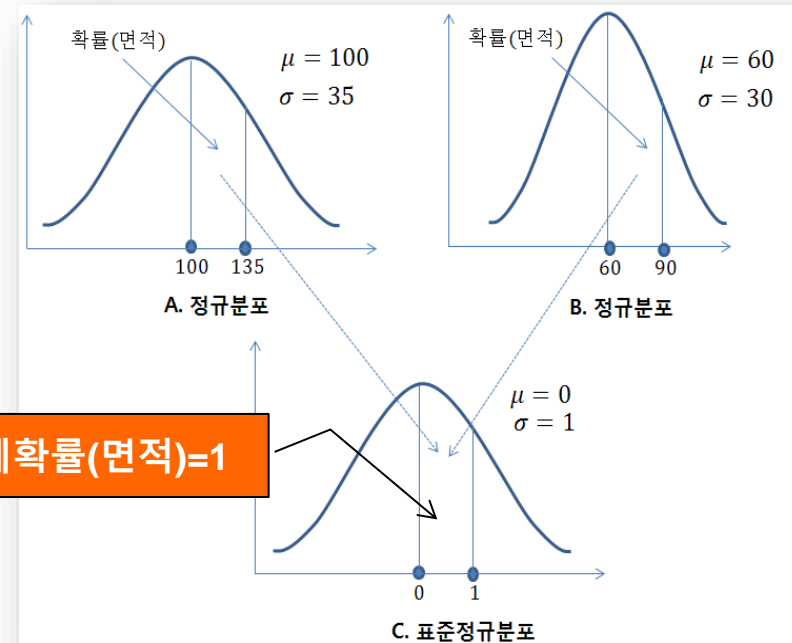
표준정규분포

● 표준정규분포(Standard Normal Distribution)란?

- ✓ 표준정규분포 또는 Z분포
- ✓ 모든 정규분포를 평균=0, 표준편차=1 표준화(정규분포 확률 문제 단순화)
- ✓ 표준화 공식 $z = \frac{x - \mu}{\sigma} \sim N(0, 1)$
- ✓ 용도 : 동일한 척도(scale) 기준으로 가치 평가



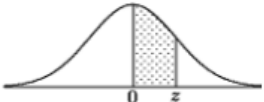
Z값 : 표준화 점수(z-score)





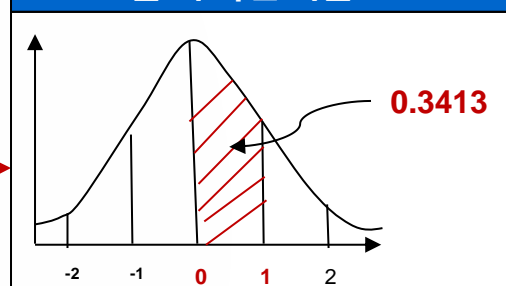
Z분포표

- Z분포표 : z값은 x축의 임계값으로 사용되며, 0~z까지의 오른쪽 확률(p)을 제공한다. 왼쪽 확률(p)이 필요한 경우 음수(-)만 붙인다.



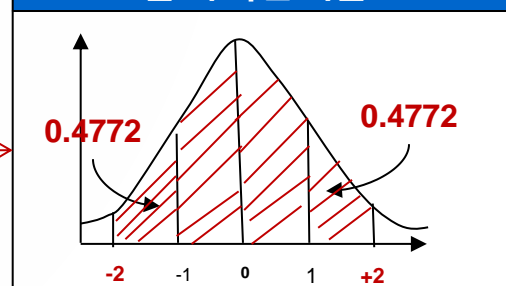
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2703	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952

z=1 일 때 곡선 확률=0.3412



$$P(0 < z < 1) = 0.3413(34.13\%)$$

z=±2 일 때 곡선 확률=0.9544



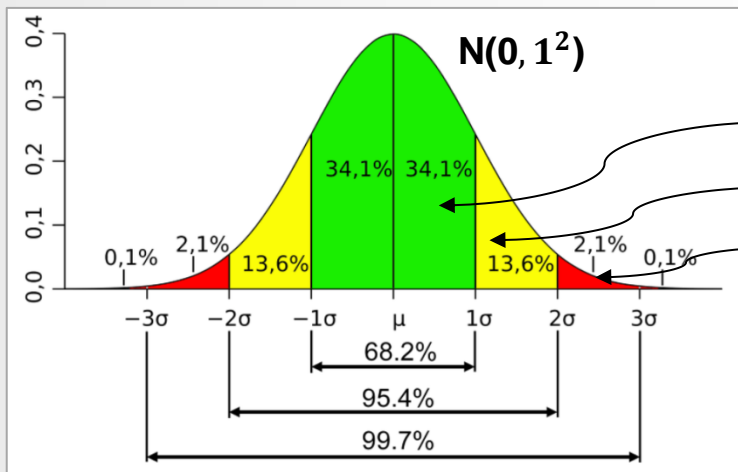
$$P(-2 < z < 2) = 0.9544(95.44\%)$$



표준정규분포 확률

● 표준정규분포 확률 : 정규분포와 동일(68-95-99.7 법칙)

- 평균 0에서 ± 1 범위 내 전체의 68.26% 존재 : $P(-1 < \mu=0 < +1) = 0.6826$
- 평균 0에서 ± 2 범위 내 전체의 95.44% 존재 : $P(-2 < \mu=0 < +2) = 0.9544$
- 평균 0에서 ± 3 범위 내 전체의 99.74% 존재 : $P(-3 < \mu=0 < +3) = 0.9974$



표준정규분포 확률표현

$$P(-1 < \mu < +1) = 0.6826(0.3413*2)$$

$$P(-2 < \mu < +2) = 0.9544(0.4772*2)$$

$$P(-3 < \mu < +3) = 0.9974(0.4987*2)$$

<Z값 확률표현 : 분포표 참고>

$$P(0 < Z < 1) = 0.3413$$

$$P(0 < Z < 2) = 0.4772$$

$$P(0 < Z < 3) = 0.4987$$

문) 95%에 해당하는 z값 확률표현은? $P(-1.96 < z < 1.96) = ?$



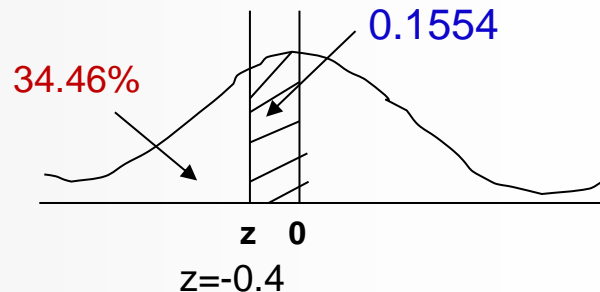
표준정규분포 적용 사례

● 정규분포 비율 계산

[사례] 성인여성(19~24세)의 키는 평균이 162cm, 표준편차가 5cm인 정규분포를 따른다고 한다. 성인 여성의 키가 160cm 이하인 비율은 얼마일까?

$$z = 160 - 162 / 5 = -0.4 \quad (p=0.1554)$$

표준정규분포에서 -0.4보다 작은 비율은 표준정규분포표에서 찾으면 0.1554이다. 키가 160cm 이하인 비율은 $0.5 - 0.1554 = 0.3446$ (**34.46%**)

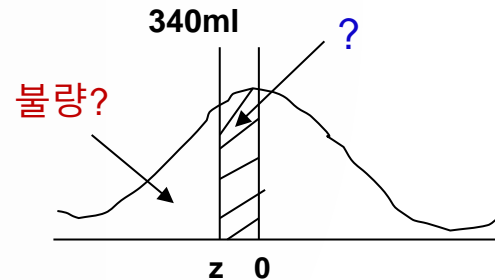




표준정규분포 적용 사례

● 정규분포 비율 계산

[문제] 스포츠 음료를 만드는 회사가 생산하는 주스는 평균 350ml, 표준편차가 3.2ml인 정규분포를 따른다고 한다. 품질 검사에서 용량이 340ml 이하이면 불량이라고 한다. 전체 제품 중에서 몇 퍼센트가 불량일까?





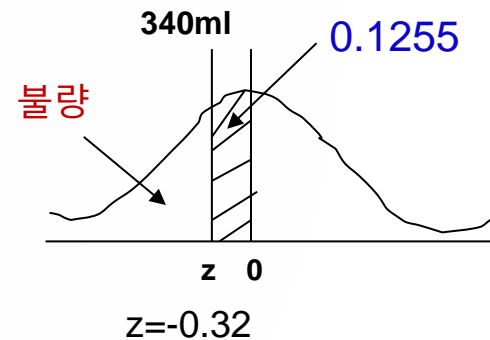
표준정규분포 적용 사례

● 정규분포 비율 계산

[문제] 스포츠 음료를 만드는 회사가 생산하는 주스는 평균 350ml, 표준편차가 3.2ml인 정규분포를 따른다고 한다. 품질 검사에서 용량이 340ml 이하이면 불량이라고 한다. 전체 제품 중에서 몇 퍼센트가 불량일까?

$$z = 340 - 350 / 3.2 = -0.32 (p=0.1255)$$

$$0.5 - 0.1255 = 37.45\%$$





5. 이항분포

● 베르누이시행(Bernoulli trial)

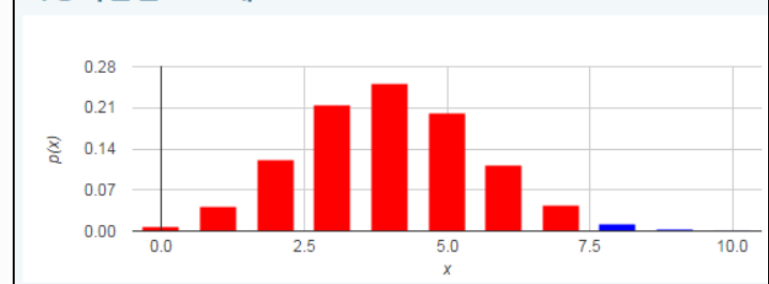
동전 던지기, 축구에서 패널티킥, 농구에서 자유투 등은 오직 두 가지 결과만 가능하다. 두 가지 실험 결과 중에서 관심 있는 사건을 성공(success), 다른 사건은 실패(failure)라고 부른다. **성공과 실패 중에서 오직 하나만 가능한 무작위 실험을 베르누이시행** 간단히 **시행(trial)**이라고 부른다. 베르누이시행의 확률분포를 베르누이분포라고 한다.

이항실험(binomial experiment)

다음 4가지 조건을 만족하는 실험을 이항실험이라고 한다.

1. 성공 확률은 p 로 표기한다.
2. 동일한 실험이 n 번 반복된다.
3. 각 시행은 '성공' 또는 '실패' 두 가지
4. 각 시행은 독립이어야 한다.

이항확률분포 그래프





이항분포

- 이항분포(binomial distribution)

성공 확률 p 인 베르누이시행을 독립적으로 n 번 반복했을 때 성공 회수 X 는 이항분포를 따르며 기호로는 다음과 같이 표기한다.

$$X \sim B(n, p)$$

k 번 성공할 확률은

$$P_r(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \dots, n$$
$$\frac{n!}{k!(n-k)!} = {}_n C_k \quad C = \text{조합(Combination)}$$

모집단의 이항분포 평균 $E(X) = np$, 분산 $V(X) = npq$, 표준편차 $\sigma(X) = \sqrt{np(1-p)}$



조합(Combination)

- **조합(Combination)** : 순서를 고려하지 않고 일부를 선택하는 경우의 수

조합의 기호는 $\binom{n}{r}$ 또는 ${}_nC_r$ 표현

의미 : n개 중에서 r개를 순서 없이 뽑는 방법의 수

공식 :
$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

예1) 친구 중 2명 뽑기

친구가 5명 있다고 가정한다. 이 중에서 2명을 뽑아서 함께 밥을 먹으려 한다. 이때 누가 먼저인지 순서는 중요하지 않고, 그냥 "함께 밥 먹을 사람 2명"만 중요

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5 \times 4}{2 \times 1} = 10 = \text{10가지}$$

예2) 팀 구성

10명 중에서 3명을 뽑아 축구팀을 만든다고 가정한다. 순서는 상관없고 그냥 누가 팀원이냐가 중요

$$\binom{10}{3} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120 = \text{120가지}$$



이항검정

● 이항검정(Binomial Test)

관측된 성공 횟수가 특정 확률 하에서 기대되는 성공 횟수와 유의하게 다른지를 검정하는 방법

[사례] A 회사에서 생산한 제품의 불량률이 5%라고 알려져 있다. 최근 생산된 100개의 제품 중 10개가 불량이었다. A 회사는 이 결과가 불량률 5%라는 주장과 다른 것인지 알고 싶어한다.

- 귀무가설 (H_0): 제품의 불량률은 5%이다. ($p = 0.05$)
- 대립가설 (H_1): 제품의 불량률은 5%가 아니다. ($p \neq 0.05$)

```
from scipy.stats import binomtest

# 관측값
x = 10      # 성공(불량품) 횟수
n = 100     # 전체 시행 횟수
p = 0.05    # 기대 확률(귀무가설 하의 성공 확률)

# 이항검정(양측 검정)
result = binomtest(k=x, n=n, p=p, alternative='two-sided')
```

결과 : p-value: 0.0341
귀무가설 기각