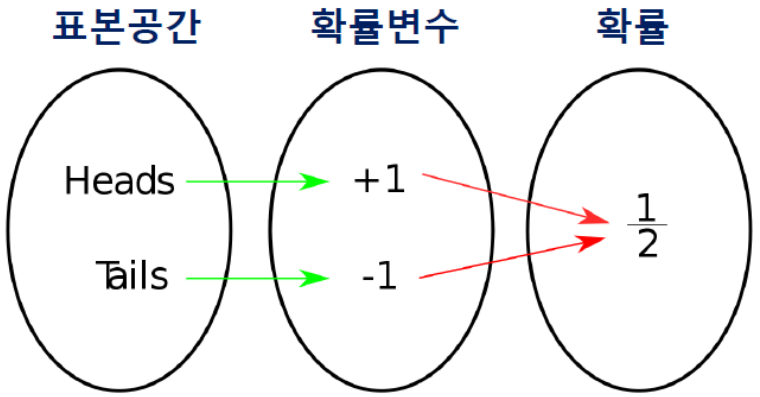
확률변수와 확률분포

표본공간(Sample space): 어떤 시행에서 일어날 수 있는 모든 결과들의 모음

확률변수(Random variable): 시행의 결과에 따라 값이 결정되는 변수

확률분포(Probability distribution): 확률변수가 특정한 값을 가질 확률을 나타내는 함수

- 동전 한 개를 던진다고 가정하면,



평균, 분산, 표준편차

데이터의 수가 n개, 데이터를 각각 $x_1, x_2, \dots, x_n$이라 한다면

평균 : $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{\sum x_n}{n}$

분산 : $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

표준편차 : $\sqrt{S^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

이산형 확률분포

이산형(discrete): 대소비교의 의미가 있는 셀 수 있는 정수 자료형

ex) 자녀수, 사고 횟수, 제품의 개수 등

이산형 변수 X의 모든 가능한 실현치 $x_1, x_2, \ldots$에 대해 확률질량 $\boldsymbol{f(x_1) = P(X = x_1), f(x_2) = P(X = x_2), \ldots}$

가 대응될 때, **X를 이산형 확률변수라 하고 $f(x_1), f(x_2), \ldots$ 를 이산형 확률분포**라 함

$f(x)$를 확률질량함수 (probability mass function) 이라 함

연속형 확률분포

확률공간 S를 가지는 연속형 변수 X에 대해

$f(x) > 0, \int_S f(x)dx = 1$.

어떤 사건 A={a<x<b} 에 대해 사건 A가 발생할 확률은

$P(X \in A) = \int_A f(x)dx = \int_a^b f(x)dx$

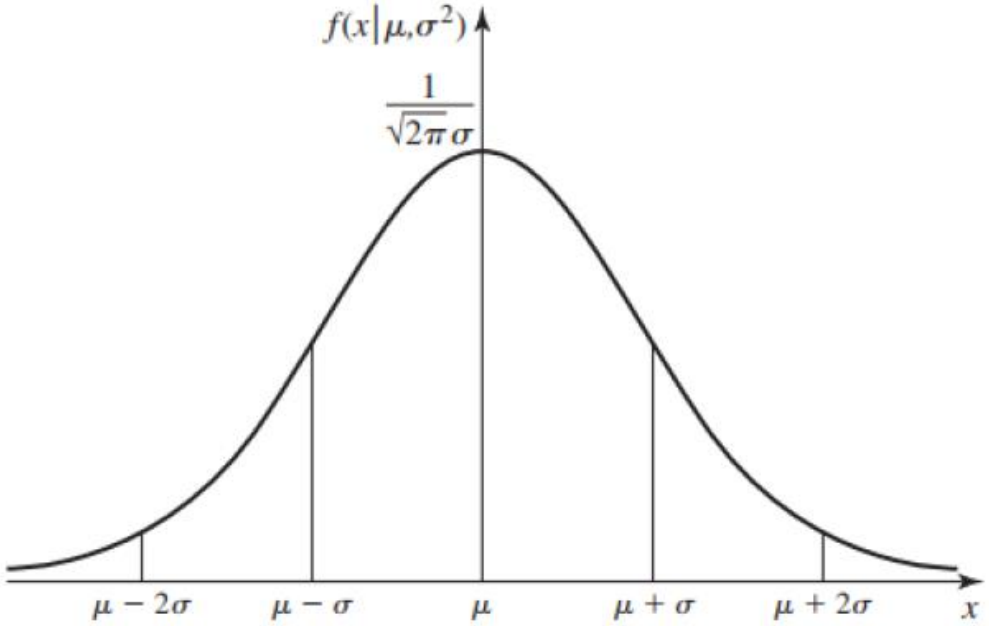
와 같을 때, X를 연속형 확률변수라 하며 함수 $f(x)$를 확률밀도함수(probability density function, pdf)라 부름

정규분포

- 독일의 수학자 가우스가 각종 물리학 실험을 수행할 때 수반되는 계측 오차에 대한 확률분포로서 가우시안분포(Gaussian distribution) 를 제시
- 연속형 자료에 대한 분포
- 많은 분야에서 확률현상을 표현하는 확률모형으로 이 분포가 자리잡음에 따라 정규분포(Normal distribution)으로 불리게 됨
- 정규분포는 주관적 입장으로 확률현상을 나타내기 위해 사용될 수 있는 하나의 확률 모형임

정규분포

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$



표준화

표준화(Standardization): 개체의 관찰에 대한 확률변수(X)와 평균(μ)의 차이를 표준편차(σ)로 나눈 값을 이 개체의 표준화 확률변수 Z로 정의

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

이렇게 구해진 표준화확률변수 Z는 본래의 확률변수 X가 따르는 확률분포가 무엇이든 평균이 0이고 표준편차가 1인 확률분포를 따르게 됨

표준정규분포

- 표준정규분포: 평균이 0이고 표준편차가1인 정규분포
- 정규분포는 평균에 따라 분포가 자리하는 위치의 중심이 결정되고 표준편차에 비례해서 분포의 폭이 커짐
- 확률변수 X가 평균이 μ이고 표준편차가 σ인 정규분포를 따를 때: $X \sim N(\mu, \sigma^2)$. $Z = \frac{X-\mu}{\sigma}$는 평균이 0이고 표준편차가 1인 정규분포를 따름: $Z \sim N(0,1)$
- 표준정규분포의 확률만 알고 있으면 어떠한 정규분포의 확률도 쉽게 계산 가능
- 표준정규분포표 이용

가설 검정

- 귀무가설(Null Hypothesis,H_0): 별다른 문제가 없는 한 나타날 것이라고 예상되는 현상에 대한 기존의 입장.
- 대립가설(Alternative Hypothesis, H_1 또는 H_A): 귀무가설 (기존의 생각)에 상반된 입장. 대안가설 또는 연구가설이라고도 함.
- 대립가설이 두 가지 방향을 주장하는 입장 (동전의 앞면이 나올 확률의 예)은 양측검정, 한 가지 방향을 주장하는 입장 (기계부품의 불량률의 예)은 단측검정이라고 함.

[귀무가설과 대립가설 예시]

- 동전의 앞면이 나올 확률
  귀무가설(H_0): 동전의 앞면이 나올 확률은 50%이다. (p=0.5)
  대립가설(H_1): 동전의 앞면이 나올 확률은 50%가 아니다. (p≠0.5)
- 기계부품의 불량률
  귀무가설(H_0): 불량률이 20%이다. (p=0.2)
  대립가설(H_1): 불량률이 20%보다 크다. (p>0.2)

가설 검정의 절차

- 단계1: 귀무가설(H_0)과 대립가설(H_1)을 수립
- 단계2: 검정을 위한 표본추출 또는 확률실험을 설계
- 단계3: 의사결정의 기준을 정함 – 귀무가설(H_0)의 기각 여부
- 귀무가설의 기각(귀무가설을 기각할 충분한 증거가 있다) -> 대립가설의 채택
- 귀무가설을 기각하지 않음(귀무가설을 기각할 충분한 증거가 없다)

  * 귀무가설을 기각하지 않는 것이 귀무가설의 채택을 의미하지는 않음

- **가설검정 절차 예시 (기계부품의 불량률)**

단계 1	H_0 : 불량률이 20%이다. $(p = 0.2)$ H_1 : 불량률이 20%보다 크다. $(p > 0.2)$
단계 2	표본으로 10개를 추출하여 불량여부를 조사
단계 3	불량품이 10개중 c개(기각치; Critical Value) 이상이면 귀무가설(H_0)을 기각

확증적 데이터 분석 (CDA)

- 확증적 데이터 분석 (Confirmatory Data Analysis, CDA)

  -> 가설을 설정한 후 수집한 데이터로 가설을 평가하고 추정하는 전통적인 분석 방법

  재현성, 유의성 검정, 신뢰구간 추정 등의 통계적 추론을 이용

1. 확증적 데이터 분석(CDA)

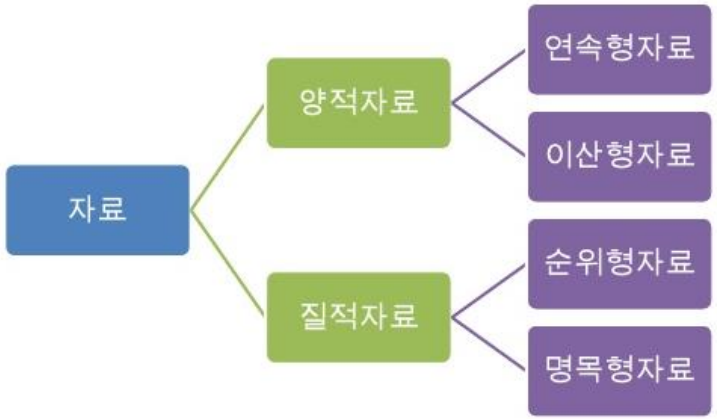
양적 자료와 질적 자료

- 양적 자료 (Quantitative Data) : 관찰 값이 수적 의미를 나타내는 자료

  -> 이산형 자료(정수값 등), 연속형 자료(무수히 많은 값)

- 질적 자료 (Qualitative Data) : 관찰 값이 수적 의미가 없이 범주만 나타내는 자료

  -> 순위형 자료(연령 등), 명목형 자료(성별 등)



일변량 양적 자료의 분석

1) 표 (Table) (구간의 빈도, 백분율)

2) 히스토그램 (Histogram)

3) 상자 그림 (Boxplot)

4) 기술통계량

- 평균 (Mean), 절사평균 (Trimmed mean), 중위수 (Median), 최빈수 (Mode), 범위 (Range), 사분위범위 (IQR: Inner Quartile Range), 분산 (Variance), 표준편차 (Standard deviation), 최소값 (Min), 최대값 (Max)

5) 분포의 모양

- 왜도, 첨도

일변량 질적 자료의 분석

1) 표 (Table) (빈도, 백분율)

2) 백분율 (Percent)

3) 막대그래프 (Bar Graph)

4) 원 그래프 (Circle Graph)

5) 기타 다른 시각화 기법들

일변량 양적 자료의 분석

1) 표 (Table) (구간의 빈도, 백분율)

2) 히스토그램 (Histogram)

3) 상자 그림 (Boxplot)

4) 기술통계량

- 평균 (Mean), 절사평균 (Trimmed mean), 중위수 (Median), 최빈수 (Mode), 범위 (Range), 사분위범위 (IQR: Inner Quartile Range), 분산 (Variance), 표준편차 (Standard deviation), 최소값 (Min), 최대값 (Max)

5) 분포의 모양

- 왜도, 첨도

가설 검정의 종류

목적	모수적 검정	비모수적 검정
정규성 검정	Shapiro-wilk 검정	
등분산성 검정	Barlett 검정 (정규성 만족 시) Levene (정규성 불만족 시)	
한 집단의 평균 비교	1 sample t-test	
독립인 두 집단의 평균 비교	Independent Sample t-test(정규성 만족 시) 윌콕슨 순위합 test (정규성 불만족 시)	Mann whitney 검정
대응표본의 차이 비교	Paired t-test	Wilcoxon 부호순위 검정
3개 이상 집단의 평균 비교	ANOVA 분석(정규성 만족 시) Kruskal - Wallis Test(정규성 불만족 시)	Kruskal-Wallis 검정
두 범주형 변수 사이의 관계	χ^2 test	

정규성 검정 (Normality test)

- 통계적 검정을 실시하기 전, 해당 표본의 모집단이 정규성을 띠는지 검정하는 방법
- Shapiro-wilk Test 주로 사용
- **귀무가설(H_0) : 해당 집단의 모집단은 정규분포를 따른다.**
- **대립가설(H_1) : 해당 집단의 모집단은 정규분포를 따르지 않는다.**
- 정규성 검정에 위배되지 않아야 대부분의 통계적 검정을 수학적으로 올바르게 사용 가능
- 하지만 모집단의 데이터가 정규성을 띠는 경우가 많지 않음. 실제 그래프를 그려 보고 분포 중심, 왜도와 첨도 등을 보고 판단해야 함
  (대략 왜도 2, 첨도 7보다 작으면 정규분포를 띤다고 봐도 무방)
- 또는 표본의 크기가 충분히 크면 ($n \geq 30$) 중심극한정리에 의해 하기 검정 사용 가능

1표본 t-검정 (1-sample t-test)

- 모집단이 정규성을 띨 경우, 모평균에 대한 t-검정
- 귀무가설(H_0) : 해당 집단의 모평균은 x일 것이다.
- 대립가설(H_1) : 해당 집단의 모평균은 x가 아닐 것이다. / x보다 (클, 작을) 것이다.

[검정 프로세스]
정규성 검정 - 정규성 만족 시 1표본 t검정

독립 2표본 검정 (Independent Two sample Test)

- 두 독립적인 집단의 모평균이 같은지, 다른지를 통계적으로 검정하는 방법
- **귀무가설(H_0) : 두 집단 간의 모평균의 차이가 없을 것이다.**
- **대립가설(H_1) : 두 집단 간의 모평균 차이는 있을 것이다 / A집단이 B집단보다 (클, 작을) 것이다.**

[검정 프로세스]

정규성 검정 - 정규성 만족 시 등분산성 검정 - 등분산성 만족 시 independent samples t-test
- 등분산성 불만족 시 welch's t-test

- 정규성 불만족 시 윌콕슨의 순위합 검정

대응 2표본 검정 (Dependent Two Sample Test)

- 동일한 집단의 사전 양적 자료와 사후 양적 자료 간에 통계적으로 유의한 차이가 있는지를 검정하는 방법
- 귀무가설(H_0) : 두 집단 간의 모평균의 차이가 없을 것이다.
- 대립가설(H_1) : 두 집단 간의 모평균 차이는 있을 것이다 / A집단이 B집단보다 (클, 작을) 것이다.

[검정 프로세스]
정규성 검정 - 정규성 만족 시 paired t-test
- 정규성 불만족 시 윌콕슨의 부호순위 검정

분산분석 (ANOVA Analysis)

- 3개 이상의 집단 간에 양적 자료에 차이가 있는지를 통계적으로 검정하는 방법
- 일원 분산분석 : 독립변인 1개, 종속변인 1개, 이원 분산분석 : 독립변인 2개, 종속변인 1개
  (종속변인 2개 이상 : MANOVA 사용)
- **귀무가설(H_0) : 독립변인에 따른 종속변인의 모평균의 차이가 없을 것이다.**
- **대립가설(H_1) : 독립변인에 따른 종속변인의 모평균의 차이가 있을 것이다.**

[검정 프로세스]
정규성 검정 - 정규성 만족 시 ANOVA analysis
- 정규성 불만족 시 Kruskal - Wallis Test

χ^2검정 (카이제곱 검정)

- 두 개의 **질적 자료** 간의 관련성이 있는지를 통계적으로 검정하는 방법

  1) 적합도 검정(goodness of fit) : 한 범주형 변수의 각 그룹별 비율이 특정 비율과 같은지 검정

  2) 동질성 검정(Test of Homogeneity) : 부모집단(subpopulation)별 분포가 동일한지 검정

  3) 독립성 검정(Test of Independence) : 두 변수의 연관관계가 유의한지 검정

	비교 대상		
적합도 검정 (goodness of fit test)	[범주형 변수] Group 1 Group 2 ... Group r	VS.	[알려진 사실] $p_1 : p_2 : \cdots : p_r$
독립성 검정 (Test of Independence)	[범주형 변수 A] Group 1 Group 2 ... Group r	VS.	[범주형 변수 B] Group 1 Group 2 ... Group r
동질성 검정 (Test of Homogeneity)	[부 모집단] Population 1 Population 2 ... Population r	VS.	[범주형 변수] Group 1 Group 2 ... Group r