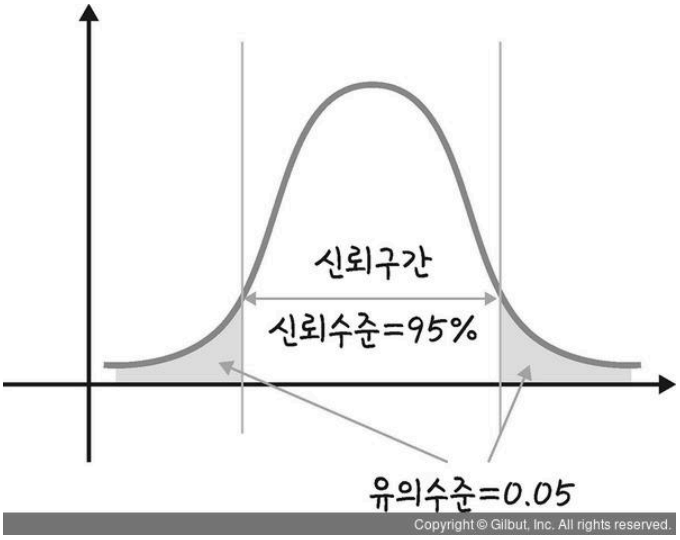
# 2회차

## 신뢰구간(Confidence Interval)

> 모집단의 평균(혹은 비율 등)이 포함될 것으로 예상되는 값의 범위 - 구간추정

- **신뢰구간** : 점추정치 ± 오차범위
- **신뢰수준** : 구간이 모수를 포함할 확률




유의수준은 가설검정 파트 설명 참고

예) 95%의 신뢰수준에서 택배가 도착하는 데 걸리는 시간은 1~1.5일입니다.

→ 해석: 95% 확률로 택배가 도착하는 데 1 ~ 1.5일 걸린다

- 신뢰구간 : 1 ~ 1.5일 ( 1.25 $\pm$ 0.25 일)
  - 점추정치 : 1.25, 오차범위: 0.25
- 신뢰수준 : 95%

### ▼ 참고) 신뢰구간 계산

#### Z분포(표준정규분포) 기반 신뢰구간

모집단 표준편차를 알고 있을 때 사용

$$$$

📎

- $\bar{X}$ : 표본평균
- $\sigma^2$: 모집단의 표준편차
- $z_{\alpha/2}$ : 신뢰수준에 따른 Z값
- $\frac{\sigma}{\sqrt{n}}$ : 표준오차

#### t분포 기반 신뢰구간

모집단 표준편차를 모를 때 사용

- $\bar{X}$ : 표본평균
- $s$ : 표본의 표준편차
- $t_{\alpha/2,df}$ : 신뢰수준에 따른 t값
- $\frac{s}{\sqrt{n}}$ : 표준오차를 근사한 값( 표본의 표준편차로 모집단의 표준편차 대체)

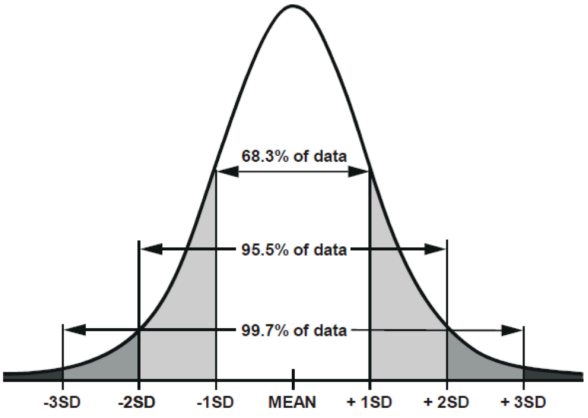
---

# 분포(Distribution)

## 정규분포(Normal Distribution)

> 종모양의 확률 분포
> - 평균에서 확률이 가장 솟아오르고, 평균을 중심으로 멀어지면서 확률이 낮아진다.

### 정규분포의 특징




1. 좌우 대칭
   - 평균 = 중앙값 = 최빈값
2. 범위(신뢰구간)
   - ±1σ 범위 → 약 68% 확률
   - ±2σ 범위 → 약 95.5% 확률
   - ±3σ 범위 → 약 99.7% 확률
3. 더하기, 빼기, 나누기를 해도 여전히 정규분포 유지 → 표준화가 가능한 이유

### 표준화(Standardization)

- 분포의 평균과 분산 값을 0과 1로 통일하는 작업
- 표준화(standard scaler) 공식: X (값) 에서 평균 $\mu$을 빼고 표준편차 $\sigma$로 나눈 값

- 표준화의 예시

> 🗣️ 은중이는 이번 시험에서 수학은 85점, 영어는 90점을 받았다.
> 은중이네 반 학생들의 성적을 고려했을 때, 은중이는 어떤 과목을 더 잘봤다고 할 수 있을까?
>
> **은중이네 반 전체 성적**
>
| 학생 | 수학 | 영어 |
|---|---|---|
| 은중 | 85 | 90 |
| 상연 | 77 | 95 |
| 재준 | 92 | 93 |
| 상학 | 70 | 92 |
| 연진 | 65 | 87 |
| 루미 | 63 | 79 |
| 세리 | 76 | 88 |
| 병헌 | 95 | 95 |
| 재석 | 80 | 77 |
| 석진 | 68 | 78 |
>
> **과목 평균과 표준편차**
>
| 과목 | 평균 | 표준편차 |
|---|---|---|
| 수학 | 77.1 | 7 |
| 영어 | 87.4 | 11 |

> 은중이는 두 과목 모두 평균보다 좋은 점수를 받았다.
>
> **은중이 점수(원점수, Z-score)**
>
| 과목 | 점수 | 표준화점수 (Z-score) |
|---|---|---|
| 수학 | 85 | 0.72 |
| 영어 | 90 | 0.37 |

> 표준화 점수를 보니 은중이는 반의 전체 수준을 고려했을 때 수학을 영어보다 잘봤다는 것을 알 수 있다.

- 표준정규분포(Z-분포, Standard Normal Distribuition, Z-Distribution)

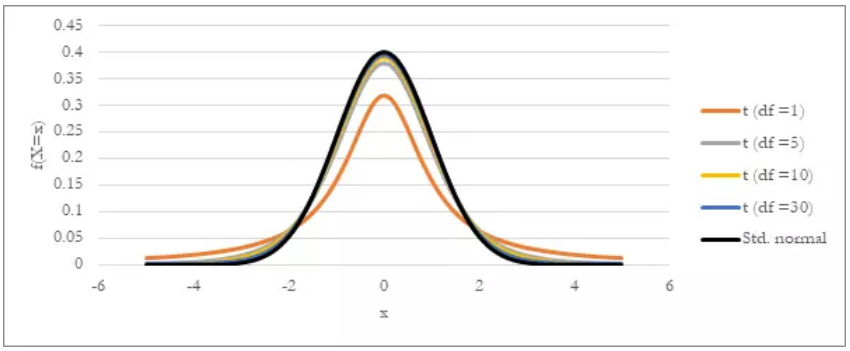
> 평균이 0이고 분산이 1인 정규분포

$$Z \sim N(0,1)$$

---

## t-분포(t-Distribution)

> **모집단의 표준편차 σ를 모를 때**, 대신 표본의 표준편차 s를 추정해서 사용하는 분포
> s의 불확실성을 반영하기 위해 정규분포보다 꼬리가 더 두꺼운 t-분포를 사용
> 특히 표본 수가 적을 때(n < 30) 사용




- 자유도(df)가 커질수록 t값은 작아진다 = t-분포는 Z 분포에 가까워진다!
  - 표본의 개수가 많아질수록, 표본표준편차 s가 σ에 더 가까워짐

- 추정의 불확실성이 줄어들면서 t-분포의 꼬리도 얇아지며 정규분포와 유사해진다.

---

### 카이제곱 분포

> **정규분포를 따르는 독립 확률변수들의 제곱합**으로 만들어지는 분포

$$$$

- $Z_i \sim N(0,1)$ : 독립적인 표준정규분포 확률변수
- k : 자유도(degrees of freedom, df)
- $\chi^2$ : 카이제곱 확률변수

## F분포

> 두 개의 **카이제곱 분포 확률변수 비율**로 정의되는 분포

$$$$

- $\chi_1^2, \chi_2^2$ : 독립적인 카이제곱 확률변수
- $df_1, df_2$ : 자유도

---

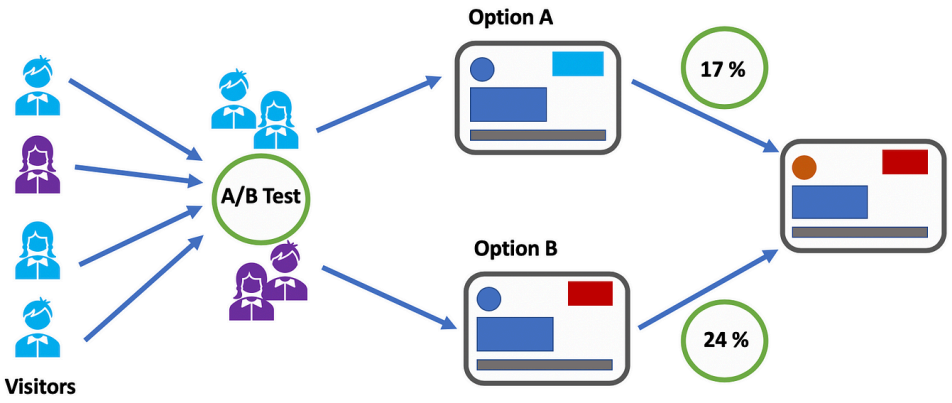
# 가설검정

> **가설검정이란 관찰된 효과가 우연에 의한 것인지 여부를 판단하는 과정을 의미합니다.**

## 가설검정의 대표 예시 : A/B TEST

- **A와 B 두 가지 처리(조건)를 비교**해서 **어떤 것이 더 효과적인지 통계적으로 검정**하는 방법

  흔히
  - A = 기존 버전 (Control) - 대조군
  - B = 새로운 버전 (Treatment) - 실험군




---

# 가설검정의 주요 개념

| 항목 | 설명 |
|---|---|
| 귀무가설($H_0$) | 밝히고자 하는 가설의 부정 : 두 그룹의 클릭률 차이는 없다 (즉, 차이 = 0) |
| 대립가설($H_1$) | 밝히고 싶은 가설 : 두 그룹의 클릭률 차이는 있다 (즉, 차이 ≠ 0) |
| 검정통계량 | 표본에서 관찰한 값이 귀무가설 아래에서 얼마나 평범한 값인지 보여주는 숫자 |
| 유의수준 α | 귀무가설($H_0$)을 기각할 기준<br>- 보통 0.05 (5%) |
| p-value | 실제 데이터에서의 차이가 우연히 나올 확률 |
| 판단 기준 | p-value < 0.05 → $H_0$ 기각 (차이 유의함) |

## 검정통계량

> 표본에서 관찰한 값이 귀무가설 아래에서 얼마나 평범한 값인지 보여주는 숫자
> **표본 데이터를 한 숫자로 요약해서, '이게 흔한 일인가, 드문 일인가?' 판단하는 척도**

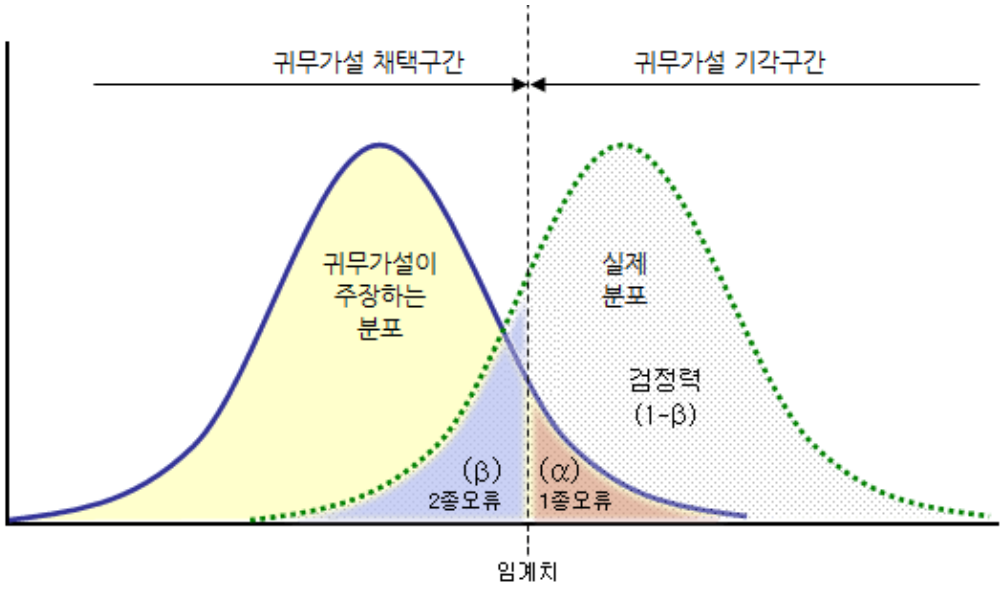
## p-value와 유의수준

> 귀무가설이 맞다고 가정했을 때, 지금과 같은 데이터가 나올 확률

- 유의수준: p-value의 수치를 통계적으로 유의한지 여길 것에 대한 기준점
  - 유의수준을 5%라 가정 → p-value가 0.05 이하이면 통계적으로 유의하다 ( = 결과가 우연히 발생한 게 아니다)
- 유의수준(α)과의 관계
  - 유의수준 α: 기준선 역할을 하는 값. 보통 0.05 (5%) 사용 (=귀무가설 채택 기준)
- p-value 에 따른 판단
  - **p-value < α → 귀무가설 기각 = 우연히 일어났을 가능성이 거의 없다 = 통계적으로 유의미하다**
  - **p-value ≥ α → 귀무가설 기각 못함 = 우연히 일어났을 가능성이 높다 = 통계적으로 유의미하지 않다.**

## 제 1종 오류와 제 2종 오류

> 가설검정에서 두 가지의 오류가 발생할 수 있다.



| | 귀무가설($H_0$) 이 참 | 귀무가설($H_0$) 이 거짓 |
|---|---|---|
| 귀무가설($H_0$) 채택 | 올바른 결정 | **제 2종 오류** |
| 귀무가설($H_0$) 기각 | **제 1종 오류** | 올바른 결정 |

## 가설검정의 기본 순서

✅ STEP1: 가설 설정 - 귀무가설과 대립가설 각각 설정

STEP2: 가설에 적합한 검정 방법 선택

STEP3: 유의수준 결정

STEP4: 검정방법에 따라서 표본의 검정통계량과 p-value 계산

STEP5: 유의수준과 p-value를 비교하여 귀무가설의 기각여부 결정

## 가설검정의 종류

분포에 따른 구분

| 검정 방식 | 검정통계량 | 관련 분포 | 활용대상 | 대상 |
|---|---|---|---|---|
| Z-검정 | $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ | 표준정규분포 | 집단 개수: 주로 2개<br>표본의 평균 비교<br>**모집단의 분산을 알 수 있는 경우** | 연속형 자료 |
| t-검정 | $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ | t분포 | 집단 개수: 주로 2개<br>표본의 평균 비교<br>**모집단의 분산을 알 수 없는 경우** | 연속형 자료 |
| 카이제곱검정 | $\chi^2 = \sum \frac{(O-E)^2}{E}$<br>$\chi^2 = \sum_{i=1}^{k} \frac{(\bar{X}_i-\mu_i)^2}{\sigma_i^2/n}$ | 카이제곱분포 | 집단 개수: 주로 2개 이상<br>**독립성 검정 : 두 범주형 변수가 독립적인지**<br>**적합도 검정 : 데이터가 특정 분포를 따르는지**<br>**동질성 검정 : 여러 집단이 동일한 분포를 따르는지** | **범주형 자료** |
| F-검정 | $F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$<br>$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{\sum_{j=1}^{k} n_j(\bar{X}j-\bar{X}\text{전체})^2}{k-1}$<br>$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2}{N-k}$ | F분포 | **집단 개수: 주로 3개 이상**<br>**두 개 이상의 그룹의 분산 비교**<br>**3개 이상의 집단 간 평균의 차이 비교**<br>회귀모델에서의 전체 유의성 검정 | 범주형 자료 |

상황에 따른 구분

| 비교 대상 | 상황 | 검정 | 관련 분포 | 전제조건 |
|---|---|---|---|---|
| **평균** | 한 집단 vs 기준값 | **단일표본 t검정** | t분포 | 정규성 |
| | 독립된 두 집단 | **이표본 t검정** | t분포 | 정규성, 등분산성 |
| | 대응된 두 집단 | **대응표본 t검정** | t분포 | 정규성 |
| | 세 집단 이상 | **ANOVA** | F분포 | 정규성, 등분산성, 독립성 |

- 현실적으로 모집단의 분산을 알 수 있는 경우가 많지 않기 때문에 Z-검정보다는 t-검정을 주로 사용합니다.

# t검정

> 1개 또는 2개의 집단 간 **평균 차이**를 비교하는 검정 방법. 정규분포를 가정.

## t-검정의 전제

- **정규성:** 정규분포에서 나온 데이터라는 전제를 가짐
- **등분산성 :** 비교 대상의 분산이 같다

  → 전제가 어긋날 경우 비모수검정(non-parametric test)을 고려해야 함

## t-검정의 종류

- **단일표본 t검정 (One-Sample t-test)**
  - 하나의 집단 평균이 특정 기준값과 다른지 비교
  - 예 : 학생들의 평균 수면시간이 7시간과 다른가?
    - 귀무가설 : 모집단의 평균은 7시간이다.
    - 대립가설 : 모집단의 평균은 7시간이 아니다.
  - 연구 맥락에 따라 의미 있는 기준값을 설정하는 것이 중요

- **이표본 t검정 (Two-Sample t-test)**
  - 독립표본 t검정과 동의어
  - 서로 독립된 두 집단의 평균 차이 비교
  - 예 : 남학생과 여학생의 평균 키가 다른가?
    - 귀무가설 : 두 집단의 평균은 같다.
    - 대립가설 : 두 집단의 평균은 다르다.
  - 전제 조건 : 정규성, 등분산성
    - 정규성이 어긋날 경우 → **Mann-Whitney U검정** 사용 (비모수 대안)
    - 등분산성이 아닐 경우 → **Welch t검정**을 사용 (정규성 가정은 여전하지만 분산은 달라져도 괜찮음)

- **대응표본 t검정 (Paired t-test)**
  - 같은 집단에서 전과 후를 비교하거나, 쌍을 이룬 데이터 비교
  - 예 : 약 복용 전후의 혈압 차이 → 비교 대상이 같은 사람, 같은 특성
  - 두 시점의 차이값(후 - 전) 자체가 정규성을 가져야 함
    - 정규성 어긋날 경우 : **Wilcoxon signed-rank 검정** 사용

  ▼ 대응 여부 구분하기

| 구분 | 독립 t검정 | 대응 t검정 |
| --- | --- | --- |
| 데이터 구조 | 두 집단이 전혀 다른 사람들 | 같은 사람의 전/후 변화 |
| 예시 | 실험군 vs 대조군 | 복용 전 vs 복용 후 |
| 검정 이름 | 이표본 t검정 | 대응표본 t검정 |

## 정규성과 등분산성 확인

📌
- 정규성 : 표본이 정규분포를 따르는 모집단에서 나왔다고 가정
- 등분산성 : 두 집단의 분산이 동일하다고 가정

### 정규성 검정 방법

| 방법 | 설명 | 사용 시기 |
|---|---|---|
| Q-Q플랏 | 정규분포와 데이터의 분위수를 비교하는 시각적 도구 | 탐색적 단계, 직관적 확인 |
| 샤피로-윌크 검정 | **귀무가설: "정규분포이다"** ($p < 0.05$면 정규성 기각) | 소표본일 때 효과적 |
| Kolmogorov-Smirnov 검정 (KS 검정) | 이론적 분포(정규분포 등)와 데이터 분포의 차이 검정 | 대체로 샤피로보다 덜 민감 |
| 히스토그램 확인 | 데이터 분포의 대칭성과 종모양을 시각적으로 확인 | 보조 자료로 사용 |

⇒ $p > 0.05$라도 정규성을 확정할 수는 없음 : 단지 정규분포가 아닌 것 같지 않다~ 수준의 판단

### 등분산성 검정

| 방법 | 설명 | 사용 시기 |
|---|---|---|
| Levene 검정 | 귀무가설: "두 집단의 분산은 같다" | 가장 널리 쓰이며 정규성 민감도 낮음 |
| F-검정 | 두 집단 분산이 같은지를 검정 (정규성 민감) | 정규성에 민감해 실제로는 잘 사용되지 않음 |
| Bartlett 검정 | 세 집단 이상에서 분산 동질성 검정 | 2개 이상 집단의 분산이 동일한지 검정 (정규성 가정 강함) |

⇒ 등분산성이 기각되면 **Welch t검정 또는 Welch ANOVA** 사용

### 실무에서의 판단 흐름

| 전제 조건 결과 | $n \leq 30$ | $n > 30$ |
|---|---|---|
| 정규성 만족 | t검정 | t검정 |
| 정규성 불만족 | 비모수 검정 권장 | t검정 가능 (분포 확인 후) |
| 등분산성 불만족 | Welch t검정 | Welch t검정 |

## 통계 실습

✅
1. 신뢰구간 구하기
2. t검정 실습
    a. 단일표본
    b. 독립표본(2표본)
    c. 대응표본

## 다음 수업 안내

일시 : 10/1(수) 10시

수업내용

1. ANOVA 검정, 다중검정
2. 카이제곱 검정 (적합성 검정, 독립성 검정)
3. 상관분석 (상관계수)