t-검정

숙명여자대학교 경영학부 오중산

집단간 평균비교

- 세 가지 집단간 평균비교 방법
  - ◆ 모표준편차를 알기 어렵기 때문에 *z*-검정은 한계가 있음
  - ◆ *t*-검정의 유형 구분
    - 독립표본 *t*-검정 vs. 대응표본 *t*-검정
  - ◆ ANOVA의 유형 구분
    - One-way ANOVA vs. Two-way ANOVA
    - ANOVA vs. MANOVA

구분	*z*-검정	*t*-검정	ANOVA
확률변수	모집단에서 정규분포를 띠어야 함 (모를 경우 표본크기 30개 이상)	모집단에서 **정규분포**를 띠어야 함	
모표준편차	**알아야 함**	모름	몰라도 됨(무관함)
모분산 조건	해당사항 없음	등분산 혹은 **이분산**	**등분산 조건 만족**해야 함
표본크기	가급적 30개 이상	무관함(**30개 미만**도 가능)	30개 이상
비교대상 집단	2개		2개 이상(보통 **3개 이상**)

독립표본 t-검정

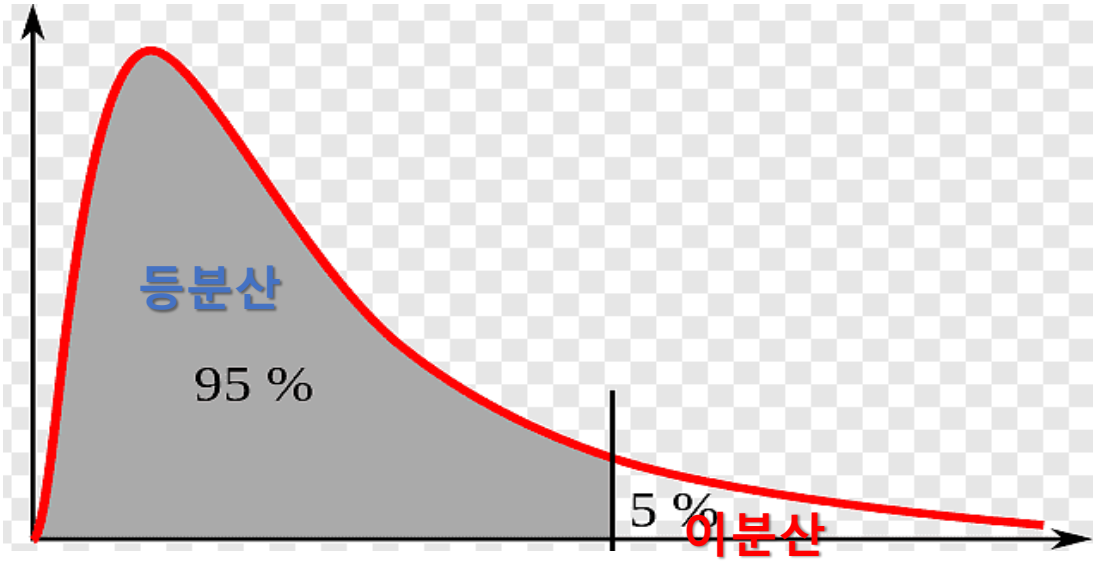
- 독립표본 t-검정 정의
  - 서로 다른 두 모집단을 대상으로 모평균 차이 유무 검정
  - 독립변수(IV)와 종속변수(DV)
    - 독립변수는 집단을 구분하는 변수로 범주형 척도로 측정
      - 집단은 두 개로 구분되어야 하므로, 만약 세 개 이상인 경우 두 개로 재분류
      - 회귀분석의 경우 독립변수와 종속변수가 모두 정량적 변수이므로 더 발전된 분석방법
    - 종속변수는 모평균 차이 비교의 대상이 되는 변수로 정량적 변수

독립표본 t-검정

- 두 가지 가설과 정규성 조건
  - 두 가지 가설
    - $H_0: \mu_1 - \mu_2 = 0$ & $H_a: \mu_1 - \mu_2 \neq 0$
  - 정규성 조건
    - t-검정을 위해서는 두 확률변수가 모두 정규성 조건을 만족해야 함
      - 두 개 표본의 표본크기가 모두 30개 이상이면 상관없지만, 30개 미만인 경우에는 t-검정을 시해하기에 앞서 사전에 정규성 조건을 확인해야 함
      - 두 확률변수가 정규분포를 띠면, 두 표본평균 각각의 표본분포도 정규성 조건 만족
      - 따라서 $\bar{X}_1 - \bar{X}_2$ 도 정규분포를 띰

독립표본 t-검정

- 등분산 검정
  - t-검정에 앞서 두 모분산이 같은지 확인해야 함
    - $H_0: \sigma_1^2 - \sigma_2^2 = 0$ & $H_a: \sigma_1^2 - \sigma_2^2 \neq 0$
  - F-통계량 활용 검정
    - $F = \frac{\hat{\sigma}_1^2(\text{큰 값})}{\hat{\sigma}_2^2(\text{작은 값})}$, $df = n - 1$
    - F-통계량이 1에 가까워야 등분산 조건 만족(H_0 채택)



독립표본 t-검정

- 등분산 가정 독립표본 t-검정
  - H_0: ${\sigma_1}^2 - {\sigma_2}^2 = 0$ 조건 만족
  - 아래와 같은 절차에 따라 t값을 구한 후, 양측검정

$\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2 \sim N(\mu_1 - \mu_2 (= 0), {\sigma^2}_{\bar{X}_1 - \bar{X}_2} (= {s_\mathrm{p}}^2 (\frac{1}{n_1} + \frac{1}{n_2})))$ $\quad s_\mathrm{p} = \sqrt{\{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2\} \div df}$ $(df = n_1 + n_2 - 2)$

$\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2 \sim N(0, (\sqrt{\frac{{s_1}^2 + {s_2}^2}{n-1}})^2)$ (만약, $n_1 = n_2 = n$) $\quad t = \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$

독립표본 t-검정

- 이분산 가정 독립표본 t-검정
  - $H_a: \sigma_1^2 - \sigma_2^2 \neq 0$ 조건 만족
  - Welch 검정을 실시하며, 아주 엄밀하게는 t값이 아니므로 t'으로 표기
  - 아래와 같은 절차에 따라 t' 값을 구한 후, 양측검정

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} \qquad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)} \qquad df = \frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{\hat{s}_1^4}{n_1^2(n_1 - 1)} + \frac{\hat{s}_2^4}{n_2^2(n_2 - 1)}}$$

독립표본 t-검정

- t-검정 절차
  - 1단계: 가설수립
  - 2단계: 집단간 데이터프레임 생성
  - 3단계: 정규성 조건 확인
  - 4단계: 등분산성 조건 확인
  - 5단계: 독립표본 t-검정 실시 및 가설검정

독립표본 *t*-검정

- *t*-검정 실습: 준비단계
  - ◆ 데이터 소개(ttest.csv)
    - E-commerce 업체에서의 고객 주문 관련 데이터
  - ◆ 데이터 프레임 만들기
    - 변수 중에서 name의 측정값 중에 특수문자(®)가 존재하므로 read.csv로 불러와 데이트 프레임 형성

변수명	변수 설명
priority	배송 우선순위
quantity	주문 물량
sales	판매금액
shipping	배송방법
price	단가
cost	주문처리비용
customer	고객유형
category	물품유형
name	물품명
container	포장크기(유형)
margin	순이익

독립표본 *t*-검정

- *t*-검정 실습: 준비단계
  - 데이터 전처리
    - 척도 변경하기
      - 문자형 척도로 측정된 변수 중에서 name을 제외한 5개의 척도를 범주형으로 변경
    - 범주형 척도 변수에 대한 빈도수 확인
    - 변수 위치 조정: 범주형/수치형/문자형 척도 측정 변수 순서로 정리
    - 이상치 검토 및 빈도수 확인
      - 수치형 척도로 측정된 5개 변수에 대해 표본평균 ± 2표본표준편차를 상/하한으로 설정
      - 5개 변수 측정값 각각에 대해 상한값을 넘어서는 이상치 빈도수 확인
    - 이상치를 제외한 데이터 프레임 생성

독립표본 *t*-검정

- *t*-검정 실습
  - 1단계: 가설수립
    - 독립변수(customer)와 종속변수(sales) 설정
      - 네 가지 customer 유형에 따른 sales 평균 비교
    - 가설수립을 위한 문제제기: “Home Office(HO) 판매금액 평균과 Consumer(CS) 판매금액 평균은 서로 동일한가?”
    - 가설수립
      - $H_0: \mu_{HO} - \mu_{CS} = 0$ & $H_a: \mu_{HO} - \mu_{CS} \neq 0$
  - 2단계: 집단별로 데이터 프레임 만들기
    - HO와 CS로 구분된 서브 데이터 프레임 생성

독립표본 *t*-검정

- *t*-검정 실습
  - 3단계: 정규성 조건 검토
    - 두 개 서브 데이터 프레임 각각에 대해 sales 관련 히스토그램 그리기
      - 히스토그램 형태를 통해 시각적으로 정규성 검토
    - shapiro.test 함수를 이용하여 sales의 정규성에 대한 통계적 검정
      - *p*-value가 유의하지 않아야 정규성 조건 만족
    - 정규성 조건 만족하지 못할 경우 대응 방안
      - 두 개 표본의 크기가 모두 크기 때문에 정규성 조건을 만족하지 못하더라도 표본평균과 표본평균 차이는 정규성 조건 만족
      - 정규성 조건에 조금 더 부합하도록 종속변수를 변환하되, 오른편으로 꼬리가 길면 자연로그 변환

독립표본 *t*-검정

- *t*-검정 실습
  - 4단계: 등분산성 조건 검토
    - var.test 함수를 이용한 등분산성 검토
      - *p*-value가 유의하지 않아야 등분산성 조건 만족
  - 5단계: 독립표본 *t*-검정 실시 및 가설검정
    - t.test 함수를 이용하여 4단계에서 등분산성 조건을 만족하면 등분산 가정 *t*-검정을 실시하고, 만족하지 못하면 이분산 가정 *t*-검정 실시
    - *p*-value가 유의하면 대립가설 채택
- *t*-검정 실습 추가
  - category에서 Technology와 Furniture 간에 sales 평균 차이 존재 여부 확인

대응표본 *t*-검정

- 대응표본(paired sample) *t*-검정이란?
  - 하나의 표본에서 서로 다른 종속변수 모평균 차이 여부 비교하는 통계분석
    - 표본이 하나이므로 집단을 구분할 필요가 없고, 독립변수 없음
  - 대응표본 *t*-검정 예시
    - (사건 없음) 동일한 소비자 집단이 한 달 동안 품목유형A를 구매한 금액평균과 품목유형B를 구매한 금액평균간의 차이 유무
    - (사건 있음) 직원들에 대해 업무몰입에 대한 동기부여 교육을 시행하기 전후의 평균업무시간 차이 유무

대응표본 *t*-검정

- 대응표본 *t*-검정 절차
  - STEP1: 가설수립
    - H_0: μ_d (두 모평균의 차이) = 0 & H_a: $\mu_d \neq 0$
    - 두 종속변수(확률변수)가 모두 정규분포를 띠거나, 표본크기가 최소 30개 이상이면 표본평균 차이는 아래와 같은 정규분포를 띰
    - s_d는 차이(X_{11}-X_{12})의 표준편차이고, 자유도는 n-1

$$\overline{X}_{11} - \overline{X}_{12} \sim N(\mu_d\ (= \mu_{11} - \mu_{12} = 0),\ \sigma^2_{\bar{X}_{11} - \bar{X}_{12}}\ (= \frac{s_d^2}{n}))$$

$$t = \frac{\overline{X}_{11} - \overline{X}_{12}}{\sigma_{\bar{X}_{11} - \bar{X}_{12}}}$$

대응표본 *t*-검정

- 대응표본 *t*-검정 절차
  - STEP2: 차이 변수 만들기
    - 두 종속변수 차이에 대한 변수(d) 생성
  - STEP3: d에 대한 정규성 검토
    - 표본이 하나이므로 등분산 조건은 확인하지 않음
  - STEP4: 대응표본 *t*-검정
    - t.test 함수에서 paired = TRUE 조건 추가