

Chapter 7. 두 모집단에 대한 비교

일표본 t-test

❖ 개요

- 모집단의 평균(μ)값을 일반적인 이해와는 다르게 연구자가 생각하는 경우 사용하는 것
- 단일 모집단 평균검증에는 모집단의 구성요소들이 정규분포를 이룬다는 가정하에 t-test를 사용
- 표본의 크기가 크면($n \geq 30$) 중심극한 정리에 따라 Z-test를 사용할 수 있으나 t-test를 사용하는 것이 보다 정확함

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \quad (df = n - 1)$$

\bar{X} : 표본의 평균값

μ_0 : 귀무가설로 설정된 모집단의 평균값

S : 표본의 표준편차

s / \sqrt{n} : \bar{X} 의 표준오차

일표본 t-test

❖ t-test 프로시저

```
PROC TTEST DATA=data-set HO=n;  
VAR variable;
```

❖ 예제

- 서비스 만족도의 모평균이 70이상인지를 알아보자.
- 귀무가설: 서비스만족도의 평균은 70이다.
- 대립가설: 서비스만족도의 평균은 70보다 크다.
- 데이터

서비스 만족도				
75	63	49	86	53
80	70	72	81	80
69	76	85	95	66
77	77	63	58	74
68	90	82	59	60

일표본 t-test

일표본 t-test

```
DATA one;  
INPUT CS@@;  
CARDS;  
75 63 49 86 53 80 70 72 81 80 69 76 85 95 66 77 77 63 58 74 68 90 82 59 60  
;  
RUN;
```

```
PROC TTEST DATA=one H0=70;  
VAR CS;  
RUN;
```

```
PROC UNIVARIATE DATA=one MU0=70 ALPHA=0.05 CIBASIC;  
VAR CS;  
RUN;
```

독립표본 t-test

❖ 개요

- 두 모집단이 정규분포를 이루고 분산이 같다($\sigma_1 = \sigma_2$)는 가정하에 t-test를 사용
- 표본의 크기가 크면($n_1 \geq 30, n_2 \geq 30$) 중심극한 정리에 따라 Z-test를 사용할 수 있으나, t-test를 사용하는 것이 보다 엄격하다는 측면에서 바람직함

$$t = \frac{\bar{X}_1 - \bar{X}_2 - D_0}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (df = n_1 + n_2 - 2)$$

\bar{X}_1 : 표본1의 평균값 \bar{X}_2 : 표본2의 평균값

μ_0 : 귀무가설로 설정된 모집단의 평균값

$$s = \sqrt{\frac{\sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_i - \bar{X}_2)^2}{n_1 + n_2 - 2}} \quad : \text{두 모집단을 결합했을 때의 결합 표준편차}(\sigma) \text{의 추정치}$$

$$s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad : \bar{X}_1 - \bar{X}_2 \text{의 표준오차}$$

독립표본 t-test

❖ t-test 프로시저

```
PROC TTEST DATA=data-set HO=n;  
CLASS variable(group);  
VAR variable(numeric);
```

❖ 예제

- 16명의 대학생들을 8명씩 랜덤 추출하여 두 집단으로 나누고 한 집단에는 기존의 강의방식(GROUP=1)을, 다른 집단에는 새로운 강의방식(GROUP=2)을 적용한 후 시험을 통해서 얻은 독해력성적(SCORE)이 차이가 있는지 알아보려고 한다.
- 귀무가설: 독해력성적에 차이가 없다.
- 대립가설: 독해력성적에 차이가 있다.
- 데이터

독해력 성적

GROUP1	65	70	76	63	72	71	68	68
GROUP2	75	80	72	77	69	81	71	78

독립표본 t-test

독립표본 t-test

```
DATA two;  
INPUT group score @@;  
CARDS;  
1 65 1 70 1 76 1 63 1 72 1 71 1 68 1 68  
2 75 2 80 2 72 2 77 2 69 2 81 2 71 2 78  
;  
RUN;  
  
PROC TTEST DATA=two COCHRAN;  
CLASS group;  
VAR score;  
RUN;
```

대응표본 t-test

❖ 개요

- 표본의 값들이 짝(pair)을 이루고 있으며, 짝을 이룬 값들을 비교하는 경우
 - 짝을 이룬 값들은 서로 독립적이지 않고, 모집단은 두 개가 아닌 하나가 됨
- t-test를 사용하지만 검증통계량과 자유도가 독립표본인 경우와 다름

$$t = \frac{\bar{d} - D_0}{S_d / \sqrt{n}} \quad (df = n - 1)$$

\bar{d} : 각 표본요소의 값들의 차이의 평균값

D_0 : 귀무가설로 설정된 차이의 평균

$S_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$: 표본요소들의 차이 값들의 표준편차

S_d / \sqrt{n} : \bar{d} 의 표준오차

대응표본 t-test

❖ t-test 프로시저

```
PROC TTEST DATA=data-set HO=n;  
PAIRED variable-1*variable-2;
```

❖ 예제

- 15명에게 컴퓨터 교육을 실시하기 이전의 성적(pretest)과 이후의 성적(posttest)을 얻었다. 시행 이후의 성적이 상승했다고 할 수 있는가?
- 귀무가설: 성적에 차이가 없다.
- 대립가설: 성적이 상승 했다.
- 데이터

컴퓨터 교육 성적

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Pre	80	73	70	60	88	84	65	37	91	98	52	78	40	79	59
post	82	71	95	69	100	71	75	60	95	99	65	83	60	86	63

대응표본 t-test

대응표본 t-test

```
DATA three;  
INPUT id pretest posttest @ @;  
CARDS;  
1 80 82 2 73 71 3 70 95 4 60 69 5 88 00  
6 84 71 7 65 75 8 37 60 9 91 95 10 98 99  
11 52 65 12 78 83 13 40 60 14 79 86 15 59 62  
;  
RUN;  
  
PROC TTEST DATA=three;  
PAIRED pretest*posttest;  
RUN;
```

단일모집단 비율검증(Z)

❖ 개요

- 단일모집단 비율검증을 위해서는 기본적으로 이항분포를 사용
- 표본의 크기가 큰 경우($n \geq 30$) 중심극한 정리에 따라 비율의 표본분포가 정규분포에 가까워져 Z-test를 사용

$$Z = \frac{\hat{p} - p_0}{SE_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

\hat{p} : 비율추정치로서 표본의 비율값

p_0 : 귀무가설로 설정된 모집단의 비율값

$$q_0 = 1 - p_0$$

$SE_{\hat{p}}$: \hat{p} 의 표준오차

단일모집단 비율검증(Z)

❖ FREQ 프로시저

```
PROC FREQ DATA=data-set ORDER=DATA;  
EXACT BINOMIAL;  
TABLE variable/BINOMIAL(P=p) ALPHA=alpha;
```

❖ 예제

- 상품에 대한 평가에 대하여 GOOD이 346명, POOR이 54명이 나왔다. 이에 대하여 POOR라고 생각하는 사람의 비율이 0.15인지 알아보자.
- 귀무가설: 비율이 0.15이다.
- 대립가설: 비율이 0.15가 아니다.

단일모집단 비율검증(Z)

단일모집단 비율검증

```
DATA four;  
INPUT state$ count;  
CARDS;  
POOR 54  
GOOD 346  
;  
RUN;  
  
PROC FREQ DATA=four ORDER=DATA;  
WEIGHT count;  
EXACT BINOMIAL;  
TABLE state / BINOMIAL(P=0.15) ALPHA=0.05;  
RUN;
```

두 모집단 비율차이 검증(Z)

❖ 개요

- 두 모집단의 비율차이 검증을 위해서는 기본적으로 이항분포를 사용
- 표본의 크기가 큰 경우($n_1 \geq 30, n_2 \geq 30$) 중심극한 정리에 따라 비율차이의 표본분포는 정규분포에 가까워져 Z-test를 사용

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}}$$

\hat{p}_1 : 비율추정치로서 표본 1의 비율값 \hat{p}_2 : 비율추정치로서 표본 2의 비율값

p_1 : 모집단 1의 비율값

p_2 : 모집단 2의 비율값

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\hat{q} = 1 - \hat{p}$$

$\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}$: $\hat{p}_1 - \hat{p}_2$ 의 표준오차

두 모집단 비율차이 검증(Z)

❖ FREQ 프로시저

```
PROC FREQ DATA=data-set ORDER=DATA;  
TABLE variable-1*variable-2 / CHISQ FISHER  
NOPERCENT NOCOL;
```

❖ 예제

- 현 정부에 대한 지지율이 성인 남녀별로 차이가 있는가를 알아보기 위해 조사를 해 본 결과, 임의로 뽑은 250명의 성인 남자 중 110명이 현 정부를 지지하였고, 200명의 성인 여자 중 104명이 현 정부를 지지하는 것으로 나타났다. 성인 남녀별로 현 정부에 대한 지지율이 차이가 있는지를 알아보자.
- 귀무가설: 남녀의 비율이 같다.
- 대립가설: 남녀의 비율이 차이가 있다.

두 모집단 비율차이 검증(Z)

두 모집단 비율차이 검증

```
DATA five;  
INPUT gender$ yesno$ count @@;  
CARDS;  
MALE YES 110 MALE NO 140  
FEMALE YES 104 FEMALE NO 96  
;  
RUN;  
  
PROC FREQ DATA=five ORDER=DATA;  
WEIGHT count;  
TABLE gender*yesno / CHISQ FISHER NOPERCENT NOCOL;  
RUN;
```


쌍 관측에 의한 두 모비율의 비교: McNemar 검정

❖ 개요

- 대응 표본으로부터 두 표본비율을 얻은 경우
- $n_{01} \sim B(n_{01} + n_{10}, 1/2)$ 이용

$$Z = \frac{n_{01} - (n_{01} + n_{10})/2}{\sqrt{(n_{01} + n_{10})/4}} = \frac{n_{01} - n_{10}}{\sqrt{n_{01} + n_{10}}}$$

- Z^2 은 자유도가 1인 카이제곱분포를 따름
- Z^2 을 McNemar의 카이제곱 통계량이라고 함

		B		
		예	아니오	합계
A	예	n_{11}	n_{10}	n_a
	아니오	n_{01}	n_{00}	$1 - n_a$
		합계	$1 - n_b$	n

쌍 관측에 의한 두 모비율의 비교: McNemar 검정

❖ FREQ 프로시저

```
PROC FREQ DATA=data-set ORDER=DATA;  
EXACT BINOMIAL;  
TABLE variable/BINOMIAL(P=p) ALPHA=alpha;
```

❖ 예제

- 상품에 대한 평가에 대하여 GOOD이 346명, POOR이 54명이 나왔다. 이에 대하여 POOR라고 생각하는 사람의 비율이 0.15인지 알아보자.
- 귀무가설: 비율이 0.15이다.
- 대립가설: 비율이 0.15가 아니다.

쌍 관측에 의한 두 모비율의 비교: McNemar 검정

McNemar 검정

```
DATA six;  
INPUT pre$ post$ count @@;  
CARDS;  
YES NO 63 YES NO 4  
NO YES 21 NO NO 12  
;  
RUN;  
  
PROC FREQ DATA=six ORDER=DATA;  
WEIGHT count;  
EXACT MCNEM;  
TABLE pre*post / NOCOL;  
RUN;
```

연습문제 1

- ❖ 한 자동차 회사에서는 새로 생산된 자동차의 연비가 평균적으로 리터당 20km라고 한다. 이 회사의 주장이 정당한지를 검정하기 위해 자동차 20대를 임의로 선택하여 조사해 본 결과 다음과 같은 데이터를 얻었다. 적절한 가설을 세우고 유의수준 5% 하에서 검정하여라.

21.0	22.7	25.8	20.6	18.5	21.4	19.3
17.6	22.7	20.6	17.9	18.3	24.7	23.3
24.3	21.5	20.0	19.8	22.9	19.9	

- ❖ 어느 회사의 평균 출근율이 90%라고 한다. 3일의 연휴 다음날 출근상태를 조사한 결과 200명 가운데 40명이 결근하였다. 이 때 평균출근율과 같다고 말할 수 있는지 유의수준 5% 하에서 검정하여 보아라.

연습문제 2

❖ A, B 두 생명보험에 대하여 계약파기비율이 다른지를 조사하고자 한다. 지난 1년간의 데이터는 다음과 같다. 두 보험에 대한 계약파기비율이 다르다고 할 수 있는지를 유의수준 5% 하에서 가설을 검정하여 보아라.

	총 계약자 수	계약파기자 수
A 생명보험사	2350	580
B 생명보험사	5210	1500

❖ 벼의 시험재배에서 품종 A와 품종 B의 수확량을 비교하고자 한다. 임의로 24개의 지역을 택하여 두 품종의 벼를 경작하여 다음과 같은 자료를 얻었다. A 품종이 B 품종보다 더 수확량이 많다고 할 수 있는가를 검정하여 보아라.

A	31	34	29	26	32	35	38	34	30	29	32	31
B	26	24	28	29	30	29	32	26	31	29	32	28