5강 통계추론 1

한림대학교 금융정보통계학과 심송용교수

목차

1 모평균에 대한 추론

모평균에 대한 추론

- $X_1, X_2, \cdots, X_n$ 이 정규분포 $N(\mu, \sigma^2)$에서의 확률표본이고 표본평균 $\bar{X}$와 표본분산 S^2을 각각

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

이라고 하고 자료에서 계산한 이들 값이 각각 $\bar{x}$, s^2 이면

- 모평균 μ에 대한 100(1−α)% 신뢰구간은

$$\left(\bar{x} - t_{n-1;\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;\alpha/2} \frac{s}{\sqrt{n}}\right)$$

이다. $t_{n-1;\alpha}$ 는 자유도가 $(n-1)$인 t-분포의 $(1-\alpha)$ 분위수.

모평균에 대한 추론

- 귀무가설 $H_0 : \mu = \mu_0$ 에 대해서 검정통계량은

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

이며 이 검정통계량은 귀무가설이 참일 때 자유도가 $(n-1)$인 t–분포.

- 자료에서 계산된 검정통계량의 값을 t_0라고 하면 각 대립가설에 대한 기각역과 유의확률은 다음과 같음.

대립가설	기각역	유의확률
$H_1: \mu > \mu_0$	$t_0 > t_{n-1;\alpha}$	$\Pr[T_{n-1} > t_0]$
$H_1: \mu < \mu_0$	$t_0 < -t_{n-1;\alpha}$	$\Pr[T_{n-1} < t_0]$
$H_1: \mu \neq \mu_0$	$\|t_0\| > t_{n-1;\alpha/2}$	$2\Pr[T_{n-1} > \|t_0\|]$

모평균에 대한 추론

예제 3.1. 학생 10명의 학업성취도를 알아보기 위해 시험을 치른 결과 성적은 다음과 같았다.
평균점수에 대한 95% 신뢰구간을 구하고, 평균이 70보다 큰지 유의수준 5%에서 검정을 해보아라.

63 72 73 70 77 72 74 73 69 79

```
> x <-c(63, 72, 73, 70, 77, 72, 74, 73, 69, 79)   # score1.r
> mx <- mean(x)   # 72.2
> sx <- sd(x)         # 4.391912
> lbd <- mx - qt(0.975, 9)*sx/sqrt(10)
> ubd <- mx + qt(0.975, 9)*sx/sqrt(10)
> lbd
[1] 69.05822
> ubd
[1] 75.34178
```

모평균에 대한 추론

```
> t0 <- (mx-70)/(sx/sqrt(10))
> t0
[1] 1.584051
> 1-pt(t0, 9)
[1] 0.0738213
```

- 위의 계산 결과 평균점수에 대한 대한 95% 신뢰구간은 (69.058, 75.342).
- 모평균에 대한 검정 $H_0 : \mu = 70$ 대 $H_0 : \mu > 70$ 에 대한

  검정통계량값은 1.584 이고, 이에 대한 단측검정의 유의확률값은 0.0738 임.
- 유의확률값 0.0738은 유의수준 0.05 보다 크므로 귀무가설을

  기각하지 못함. 즉, 유의수준 0.05에서 평균이 70 보다 크다고 할 수 없음.

2 독립 이표본 평균차에 대한 추론

독립 이표본 평균차에 대한 추론($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

1. $X_1, X_2, \cdots, X_m$ 이 정규분포 $N(\mu_1, \sigma_1^2)$에서의 확률표본이고
2. $Y_1, Y_2, \cdots, Y_n$ 이 정규분포 $N(\mu_2, \sigma_2^2)$에서의 확률표본이며
3. 두 확률표본이 독립이며 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 임을 가정.

$$\frac{\sum_{i=1}^m X_i}{m}, \quad S_1^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2}{m-1}, \ \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}, \quad S_2^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$$

이라고 하면 공통분산 σ^2의 추정량 S_p^2 는

$$S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2} = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2}{m+n-2}$$

로 얻으며

독립 이표본 평균차에 대한 추론($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$t = \frac{\bar{X} - \bar{Y} \ - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

의 분포는 자유도가 $(m + n - 2)$인 t–분포를 따름.

- 따라서 $\mu_1 - \mu_2$**에** 대한 100(1–α)% 신뢰구간은

$$(\bar{x} - \bar{y} - t_{m+n-2;\,\alpha/2}\ s_p\sqrt{\frac{1}{m} + \frac{1}{n}}\ ,\ \bar{x} - \bar{y} + t_{m+n-2;\,\alpha/2}\ s_p\sqrt{\frac{1}{m} + \frac{1}{n}}\ )$$

- 귀무가설 $H_0 : \mu_1 - \mu_2 = \varDelta_0$ **에 대해서 검정통**계량은

$$t = \frac{\bar{x} + \bar{y} \ - \varDelta_0}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

이며

독립 이표본 평균차에 대한 추론($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

- 자료에서 계산된 검정통계량의 값을 t_0 라고 하면 각 대립가설에 대한 기각역과 유의확률은 다음과 같음.

대립가설	기각역	유의확률
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$t_0 > t_{m+n-2;\alpha}$	$\Pr[T_{m+n-2} > t_0]$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$t_0 < -t_{m+n-2;\alpha}$	$\Pr[T_{m+n-2} < t_0]$
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$\|t_0\| > t_{m+n-2;\alpha/2}$	$2\Pr[T_{m+n-2} > \|t_0\|]$

독립 이표본 평균차에 대한 추론($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

예제 3.2. 남학생 10명과 여학생 11명의 체질량지수를 조사하였더니 다음과 같았다.
남학생의 체질량지수의 평균이 여학생보다 크다고 할 수 있는지 유의수준 5%에서 검정하여라. 체질량지수 평균의 차이에 대한 95% 신뢰구간을 구해보아라.
(두 집단의 분산이 같다고 가정).

남학생 : 21.6 20.8 17.6 20.1 20.1 21.9 20.6 19.4 21.5 26.1
여학생 : 20.6 20.4 20.2 20.2 18.0 19.8 20.9 19.7 20.3 19.7 22.7

```
> x <- c(21.6,20.8,17.6,20.1,20.1,21.9,20.6,19.4,21.5,26.1)
> y <- c(20.6,20.4,20.2,20.2,18.0,19.8,20.9,19.7,20.3,19.7,22.7)
> mx <- mean(x) ; my <- mean(y)
> sdx <- sd(x) ; sdy <- sd(y)
> sp <- sqrt( (9*sdx^2+10*sdy^2)/(10+11-2) )
> t0 <- (mx-my)/(sp*sqrt(1/10+1/11))
```

독립 이표본 평균차에 대한 추론($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

```
> t0
[1] 0.9918929
> 1-pt(t0, 19)
[1] 0.1668573
> lbd <- (mx-my) - qt(0.975, 19)*sp*sqrt(1/10+1/11)
> ubd <- (mx-my) + qt(0.975, 19)*sp*sqrt(1/10+1/11)
> lbd
[1] -0.8245246
> ubd
[1] 2.309979
```

- 결과에서 두 그룹의 평균차에 대한 검정통계량은 0.992 이고,
- 이에 대한 단측유의확률이 $0.16686 > 0.05$ 이므로 차이가 없다는 귀무가설을 기각하지 못함. 즉 남학생의 체질량지수의 평균이 여학생보다 크다고 할 수 없음.
- 평균의 차이에 대한 95% 신뢰구간은 $(-0.8245, 2.3100)$임.

독립 이표본 평균차에 대한 추론 ($\sigma_1^2 \neq \sigma_2^2$)

1. $X_1, X_2, \cdots, X_m$ 이 정규분포 $N(\mu_1, \sigma_1^2)$에서의 확률표본이고
2. $Y_1, Y_2, \cdots, Y_n$ 이 정규분포 $N(\mu_2, \sigma_2^2)$에서의 확률표본이며
3. 두 확률표본이 독립이며 $\sigma_1^2 \neq \sigma_2^2$

- 다음의 통계량

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

은 자유도가

$$d = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{\left(\frac{S_1^2}{m}\right)^2}{m-1} + \frac{\left(\frac{S_2^2}{n}\right)^2}{n-1}}$$

독립 이표본 평균차에 대한 추론 ($\sigma_1^2 \neq \sigma_2^2$)

인 t–분포를 따름. 이 자유도를 Satterswaite 자유도라고 함.

- 따라서 신뢰구간은
- $\left(\bar{x} - \bar{y} - t_{d;\alpha/2}\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}} ,\ (\bar{x} - \bar{y} + t_{d;\, \alpha/2} \sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}} \right)$

- 귀무가설 $H_0: \mu_1 - \mu_2 = \Delta_0$ 에 검정통계량은

$$t = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

이며

독립 이표본 평균차에 대한 추론 ($\sigma_1^2 \neq \sigma_2^2$)

- 각 대립가설에 대한 기각역 및 유의확률은 다음과 같음

대립가설	기각역	유의확률
$H_1: \mu_1 - \mu_2 > \Delta_0$	$t_0 > t_{d;\alpha}$	$\Pr[T_d > t_0]$
$H_1: \mu_1 - \mu_2 < \Delta_0$	$t_0 < -t_{d;\alpha}$	$\Pr[T_d < t_0]$
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	$\lvert t_0 \rvert > t_{d;\alpha/2}$	$2\Pr[T_d > \lvert t_0 \rvert]$

3 짝비교

짝비교

- $(X_1, Y_1), (X_2, Y_2), \cdots , (X_n, Y_n)$ 이 독립인 쌍이라고 하자. 이 때 X_i와 Y_i의 기댓값은 각각 μ_1와 μ_2 이며 일반적으로 X_i와 Y_i 는 독립이 아님.
- 차이 D_i가

$$D_i = X_i - Y_i$$

이면 D_i 는 모두 독립이며 정규분포 정규분포 N(μ_D, σ_D^2) 를 따름을 가정.

- $\overline{D}$ **와** S_D를 각각 D_i들의 평균과 표준편차라 하면 일표본 t–검정에서 본 것과 같이

$$t = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}$$

는 자유도가 $(n-1)$인 t–분포.

짝비교

- 귀무가설 H_0 : $\mu_1 - \mu_2 = \Delta_0$ 에 대해서 검정통계량은

$$t = \frac{\bar{D} - \Delta_0}{S_D/\sqrt{n}}$$

이며 이 검정통계량은 귀무가설이 참일 때 자유도가 $(n-1)$인 t–분포.

- 자료에서 계산된 검정통계량의 값을 t_0라고 하면 각 대립가설에 대한 기각역과 유의확률은 다음과 같음.

대립가설	기각역	유의확률
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$t_0 > t_{n-1;\alpha}$	$\Pr[T_{n-1} > t_0]$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$t_0 < -t_{n-1;\alpha}$	$\Pr[T_{n-1} < t_0]$
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$\lvert t_0 \rvert > t_{n-1;\alpha/2}$	$2\Pr[T_{n-1} > \lvert t_0 \rvert]$

짝비교

예제 3.5. 고혈압 환자의 최초 내원시의 이완기 혈압과 치료 한 달 후의 이완기 혈압을 조사하였더니 다음과 같았다. 치료 후의 이완기 혈압이 낮아졌다고 할 수 있는지 유의 수준 5%에서 검정하여라. 또 치료전후의 혈압차에 대한 95% 신뢰구간을 구하여 보아라.

환자	1	2	3	4	5	6	7	8	9	10
최초	106	107	105	113	107	112	109	112	111	114
치료후	100	92	91	82	87	96	101	96	79	96

짝비교

```
> x <- c(106,107,105,113,107,112,109,112,111,114)
> y <- c(100, 92, 91, 82, 87, 96,101, 96, 79, 96)
> dd <- x-y
> nn <- length(dd)
> t0 <- mean(dd) / ( sd(dd)/sqrt(nn) )
> t0
[1] 6.577429
> 1-pt(t0, nn-1)
[1] 5.094548e-05
> lbd <- md - qt(.975, nn-1)* sdd/sqrt(nn)  # md =mean(dd), sdd=sd(dd)
> ubd <- md + qt(.975, nn-1)* sdd/sqrt(nn)
> list(lbd, ubd)
[[1]] [1] 11.54688
[[2]] [1] 23.65312
```

4 R-언어의 평균 추론 함수

R-언어의 평균 추론 함수

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0,
         paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)
t.test(formula, data, subset, na.action, ...)
```

- x : 일표본 또는 이표본 검정에서 첫 번째 그룹의 자료를 저장한 벡터.
- y : 이표본 검정에서 두 번째 그룹의 자료가 저장된 벡터, 일표본 검정인 경우 설정하지 않아도 됨.
- alternative : 대립가설의 방향을 설정함. 기본값은 양측검정을 설정하는 "two.sided" 이며 '크다' 또는 '작다'인 경우 각각 "greater" 또는 "less"를 설정할 수 있으며 첫 글자만 사용하여도 된다. 여기서 '크다'와 '작다' 먼저 나온 x의 평균이 두 번째인 y의 평균보다 '크다' 또는 '작다'를 의미.
- mu : 귀무가설하에서의 값(또는)을 설정한다. 기본값은 0.
- paired : 논리값 True/False를 설정하며 대응표본인 경우 True를 설정한다. 기본값은 False.
- var.equal : 논리값을 설정하며 독립인 두 그룹 비교에서 등분산을 가정할지 True/False로 설정한다. 등분산을 가정하는 것이 기본값.

R–언어의 평균 추론 함수

- conf.level : 신뢰구간의 신뢰도를 설정.

참고 : 신뢰구간은 alternative가 "two.sided"일 때만 우리가 흔히 알고 있는 양측 신뢰구간을 계산하며,
alternative가 "greater" 또는 "less"인 경우 한쪽 신뢰구간만 계산함.

- formula : 독립인 두 그룹을 설정하는 방법으로 dep ~ indep 로 설정하며 dep는 종속변수,
  indep는 그룹(이 경우 두 개의 그룹)을 표시하는 변수를 설정.
- data : 사용할 데이터 프레임의 이름을 설정.
- subset : data에 설정한 데이터 프레임에서 일부의 자료만 얻고자 할 때 해당하는 조건식을 설정.
- na.action : 자료에 NA가 있을 때 NA를 어떻게 처리할지 설정한다. 설정은 NA의 값을 처리할
  함수를 설정하며, 기본값은 getOption("na.action")에 설정된 값으로 한다.
  R 기본 설치 시 설정된 값은 NA를 제외하는 것.

R-언어의 평균 추론 함수

이 함수의 개별 출력은 다음과 같이 얻을 수 있다.

- statistic : 검정통계량의 값 t_0의 값을 출력.
- parameter : 검정통계량의 자유도를 출력.
- p.value : 유의확률을 출력.
- conf.int : 신뢰구간을 출력.
- estimate : 평균의 값(일표본일 때) 또는 평균의 차이(이표본일 때)를 출력.
- null.value : 귀무가설의 평균값 또는 평균의 차이를 출력.
- alternative : 대립가설의 방향을 출력.
- method : 검정의 종류를 출력.
  ("One Sample t-test", " Two Sample t-test", "Welch Two Sample t-test", "Paired t-test").
- data.name : 데이터 프레임이 사용된 경우 데이터 프레임의 이름을 출력.

R-언어의 평균 추론 함수

예제 3.6. (일표본 t 검정) 예제 3.1.의 자료를 t.test 함수에 적용하여 보자

```
> x <-c(63, 72, 73, 70, 77, 72, 74, 73, 69, 79)
> t.test(x,alternative="greater", mu=70)
        One Sample t-test
data:  x
t = 1.5841, df = 9, p-value = 0.07382
alternative hypothesis: true mean is greater than 70
95 percent confidence interval:
 69.65409       Inf
sample estimates:
mean of x
     72.2
```

R-언어의 평균 추론 함수

- 결과에서 검정통계량의 값은 1.584, 자유도 9이며 유의확률은 0.0738로 얻어서 예제 3.1.의 결과와 일치함.
- alternative에 "greater"를 설정하였으므로 신뢰구간도 일방향 신뢰구간만 얻음.
- 일반적인 신뢰구간을 얻으려면 alternative를 생략하거나 "two"(또는 "two.sided")를 설정.

```
> x <-c(63, 72, 73, 70, 77, 72, 74, 73, 69, 79)
> t.test(x, alternative="two", mu=70)
One Sample t-test
data: x
t = 1.5841, df = 9, p-value = 0.1476
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
69.05822 75.34178
```

R-언어의 평균 추론 함수

예제 3.7. (독립 이표본 등분산인 경우) 예제 3.2.의 자료를 t.test 함수에 적용하면 다음과 같이 설정한다. 이 경우 등분산을 가정하므로 var.equal에 T로 설정하였다.

```
> x <- c(21.6,20.8,17.6,20.1,20.1,21.9,20.6,19.4,21.5,26.1)
> y <- c(20.6,20.4,20.2,20.2,18.0,19.8,20.9,19.7,20.3,19.7,22.7)
> t.test(x, y, alternative = "greater", var.equal = T)
Two Sample t-test
data: x and y
t = 0.9919, df = 19, p-value = 0.1669
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
-0.5520436 Inf
```

R-언어의 평균 추론 함수

```
sample estimates:
mean of x mean of y
20.97000 20.22727
> t.test(x, y, var.equal=T, conf=.99)$conf
[1] -1.399534 2.884989
attr(,"conf.level")
[1] 0.99
```

- 검정통계량, 유의확률이 모두 예제 3.2의 결과와 일치함.
- 신뢰구간은 alternative를 생략하여(기본값 "two.sided"가 적용됨) t.test를 설정하고, 필요한 신뢰구간만 출력으로 얻기 위해 $conf를 적용.
- 이 결과도 예제 3.2.의 신뢰구간과 일치함.

R-언어의 평균 추론 함수

예제 3.8. (독립 이표본 – 분산이 같지 않은 경우) 예제 3.2.의 자료를 적용하고 var.equal의 설정을 생략(또는 F로 설정)하여 t.test 함수를 적용. 이 경우 필요한 세 개의 출력(검정통계량, 자유도, 유의확률)을 얻기 위해 앞의 세 개만 출력.

```
> x <- c(21.6,20.8,17.6,20.1,20.1,21.9,20.6,19.4,21.5,26.1)
> y <- c(20.6,20.4,20.2,20.2,18.0,19.8,20.9,19.7,20.3,19.7,22.7)
> t.test(x,y,alternative="g")[c(1,2,3)]
$statistic
t
0.9629611
$parameter
df
13.07333
```

R-언어의 평균 추론 함수

```
$p.value
[1] 0.1765208

> # Confidence Interval
> t.test(x,y, conf=.99)$conf
[1] -1.578460 3.063915
attr(,"conf.level")
[1] 0.99
```

R-언어의 평균 추론 함수

예제 3.9. (대응표본 t 검정 경우) 예제 3.5.의 자료를 사용하여 대응표본 t-검정을 하기 위해 paired에 TRUE를 설정하면 다음과 같은 결과를 얻는다.

```
> t.test(x, y, paired=T)
Paired t-test
data: x and y
t = 6.5774, df = 9, p-value = 0.0001019
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
11.54688 23.65312
sample estimates:
mean of the differences
17.6
```

R-언어의 평균 추론 함수

```
$parameter
df
13.07333
$p.value
[1] 0.1765208
> # Confidence Interval
> t.test(x,y, conf=.99)$conf
[1] -1.578460 3.063915
attr(,"conf.level")
[1] 0.99
```

- 예제 3.5.와 일치하는 결과를 얻음

5 비율에 대한 추론

비율에 대한 추론

- n개의 독립인 베르누이 시행에서 성공의 회수 X의 분포는 이항분포 $B(n,p)$이다. 이항분포에서의 성공확률 p의 추정량 $\hat{p}$는 $\hat{p} = \frac{X}{n}$이며 $\hat{p}$ 및 X의 근사분포는 각각 $N\left(p, \frac{p(1-p)}{n}\right)$ 및 $N\left(np, np(1-p)\right)$ 임(이항분포의 정규근사).

- 따라서 모비율 p에 대한 100(1−α)% (근사) 신뢰구간은

$$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}\right)$$

- 귀무가설 $H_0 : p = p_0$ 에 대해 검정통계량은

$$z = \frac{X - np_0}{\sqrt{np_0q_0}} = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}}$$

비율에 대한 추론

- 기각역 및 유의확률은 다음과 같음.

대립가설	기각역	유의확률
$H_1 \colon p > p_0$	$z_0 > z_\alpha$	$\Pr[Z > z_0]$
$H_1 \colon p < p_0$	$z_0 < -z_\alpha$	$\Pr[Z < z_0]$
$H_1 \colon p \neq p_0$	$\lvert z_0 \rvert > z_{\alpha/2}$	$2\Pr[Z > \lvert z_0 \rvert]$

- 연속성수정이 있는 경우 신뢰구간 :

$$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\,\hat{q}}{n}} - \frac{1}{2n}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\,\hat{q}}{n}} + \frac{1}{2n}\right)$$

비율차에 대한 추론

1. $X_1, X_2, \cdots, X_m$ 이 성공확률이 p_1인 베르누이 시행에서의 확률표본이고
2. $Y_1, Y_2, \cdots, Y_n$ 이 성공확률이 p_2인 베르누이 시행에서의 확률표본이며
3. 두 확률표본이 독립인 경우

- **모비율 차이 $p_1 - p_2$ 에 대한 100(1−α)% (근사) 신뢰구간은**

$$\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}, \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}} \right)$$

- **귀무가설 $H_0 : p_1 - p_2 = 0$에 대해 검정통계량은**

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}}\sqrt{\frac{1}{m}+\frac{1}{n}}}, \quad \text{단 } \hat{p} = \frac{X+Y}{m+n}$$ (주의: 교재오류)

비율차에 대한 추론

- 기각역 및 유의확률은 다음과 같음.

대립가설	기각역	유의확률
$H_1: p_1 - p_2 > 0$	$z_0 > z_\alpha$	$\Pr[Z > z_0]$
$H_1: p_1 - p_2 < 0$	$z_0 < -z_\alpha$	$\Pr[Z < z_0]$
$H_1: p_1 - p_2 \neq 0$	$\lvert z_0 \rvert > z_{\alpha/2}$	$2\Pr[Z > \lvert z_0 \rvert]$

예제 3.11. (〈R 3.12〉)의 z0로는 다음으로 수정(교재 오류)

```
> p <- (28+8)/(35+45)
> z0 <- (p1-p2)/sqrt( p*(1-p)*(1/n1 + 1/n2))
```

비율에 대한 R-함수

```
prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

- x : 일표본 비율 추론인 경우 성공의 횟수를 저장한 변수이거나 성공횟수 및 실패횟수의 행렬 또는 표(table) 개체를 설정. 일표본 비율 추론에서 x가 행렬(matrix 함수) 또는 표(table 함수)인 경우 열의 개수가 반드시 2여야 함. 이 표본 비율에서 비율차에 대한 추론인 경우 x는 2×2 인 행렬 또는 표 개체를 설정하며 첫 번째 행은 첫 번째 그룹의 성공회수, 실패횟수를 두 번째 행은 두 번째 그룹의 성공횟수, 실패횟수를 저장한 행렬 또는 표 개체를 설정.
- n : 시행횟수 n을 설정. 만일 x가 행렬이나 표이면 n 값은 무시됨.
- p : 귀무가설하에서의 성공확률값 p_0를 설정하거나, 두 그룹의 비율비교인 경우 NULL을 설정.
- alternative : 대립가설의 방향이 '크다'(" greater "), '작다'(" less "), 또는 '같지 않다'(" two.sided ")인지 설정. 기본값은 '같지 않다'임.
- conf.level : 신뢰구간의 신뢰도를 설정.

비율에 대한 R-함수

- correct : 연속성 수정을 포함할지 설정. 기본값은 연속성 수정이 포함된다.
  단, 연속성 수정은 Newcombe(1988a)에 의한 수정을 적용.

예제 3.12. 어느 대학의 학생 80명을 조사하였더니 36명이 백팩을 매고 다닌다고 한다. 백팩을 매는 학생의 비율이 40% 인지 유의수준 5%에서 검정하고 백팩을 매는 학생의 비율에 대한 95% 신뢰구간을 구하여라.

```
> xx <- 36; nn <- 80; phat <- xx/nn ;  p0 <- 0.4
> prop.test(xx, nn, p=0.4, correct=F)
1-sample proportions test without continuity correction
data: xx out of nn, null probability 0.4
X-squared = 0.8333, df = 1, p-value = 0.3613
alternative hypothesis: true p is not equal to 0.4
95 percent confidence interval:
0.3457769 0.5588049
```

비율에 대한 R-함수

```
sample estimates:
p
0.45
```

- 일표본 비율은 행렬로 성공횟수 및 실패 회수 자료를 설정하여 추론할 수 있음

```
>  x <- matrix(c(36, 44), ncol = 2)
> prop.test(x, correct = F, p = 0.4)
  (출력 결과는 앞과 같아 생략)
```

비율에 대한 R-함수

예제 3.13. 예제 3.12.의 자료를 성별에 따라 구분하여 보았더니 다음과 같았다. 남학생과 여학생의 백팩 사용 비율이 차이가 있는지 유의수준 5%에서 검정하고 비율차에 대한 95% 신뢰구간을 구하여 보아라.

	백팩 사용	백팩 미사용	합
남학생	28	7	35
여학생	8	37	45
합	36	44	80

비율에 대한 R-함수

```
> x <- matrix(c(28,7,8,37), byrow = T, ncol = 2)
> prop.test(x, correct = F)

        2-sample test for equality of proportions without continuity
        correction

data:  x
X-squared = 30.7969, df = 1, p-value = 2.865e-08
alternative hypothesis: two.sided
95 percent confidence interval:
 0.4489043 0.7955402
sample estimates:
   prop 1    prop 2
0.8000000 0.1777778
```

다음시간 안내

통계추론(2)