R을 이용한 기초 통계 분석

Jinseog Kim
Dongguk University
jinseog.kim@gmail.com

2017-04-12

Agenda

- **기본 개념**
- **t검정(t test)**
- 분산분석(ANOVA, Analysis of Variance)
- 범주형 자료의 분석
  - 동일성 검정 (Test of homogeneity)
  - 독립성 검정 (Test of independence)
- 회귀분석(regression analysis)
  - 선형회귀분석 (linear regression analysis)
  - 로지스틱 회귀분석 (logistic regression analysis)

통계학이란

1. 데이터(자료)에 근거하여 자연 혹은 사회적 현상에 대한 과학적 추론과 합리적인 의사결정을 하도록 하는 학문
2. 통계학은
    1. 관심의 대상에 대한 자료를 수집하고
    2. 수집된 자료를 정리, 요약하며
    3. 주어진 자료를 토대로 불확실한 사실에 대한 과학적인 판단을 하도록 하는 제반 방법들

기본 용어

1. 모집단 (population): 관심의 대상이 되는 모든 개체들의 관측값의 집합
   - 유한모집단: 모집단의 원소의 개수가 유한개인 경우
   - 무한모집단: 모집단의 원소의 개수가 무한개인 경우
2. 표본 (sample): 모집단에서 실제 조사한 관측값의 집합
   - 랜덤표본 (random sample)

기술통계학과 추측통계학

- 기술통계학: 수집된 자료들의 특성을 파악하기 쉽도록 단순히 정리하거나 요약하는 분야
  - 표본 평균, 표본 분산 등
  - 도수 분포표, 분할표, 상자그림 등 ...
- 추측통계학: 표본에 내포되어 있는 정보를 이용하여 모집단의 특성을 파악, 추론하는 분야
  - 추정과 검정
  - 예측

자료의 종류

- 질적 자료 (범주형 자료) : 명목형 또는 순서형 자료
  - 명목형 (nominal): 성별 (이진형; binary), 질환명 (multinomial)
  - 순서형 (ordinal): 5점척도(best, good, normal, bad, worst), ...
- 양적 자료
  - 실수형 (real-valued)
  - 정수형 (count)
  - 비율형 (ratio, rate)

Boxplot (상자그림)

1. read data and ordering: $X_{(1)}, \ldots, X_{(n)}$
2. Q_1, Q_2, Q_3 계산:

$$Q_1 = \frac{X_{([n/4])} + X_{([n/4]+1)}}{2}, \quad Q_2 = \frac{X_{([n/2])} + X_{([n/2]+1)}}{2}$$

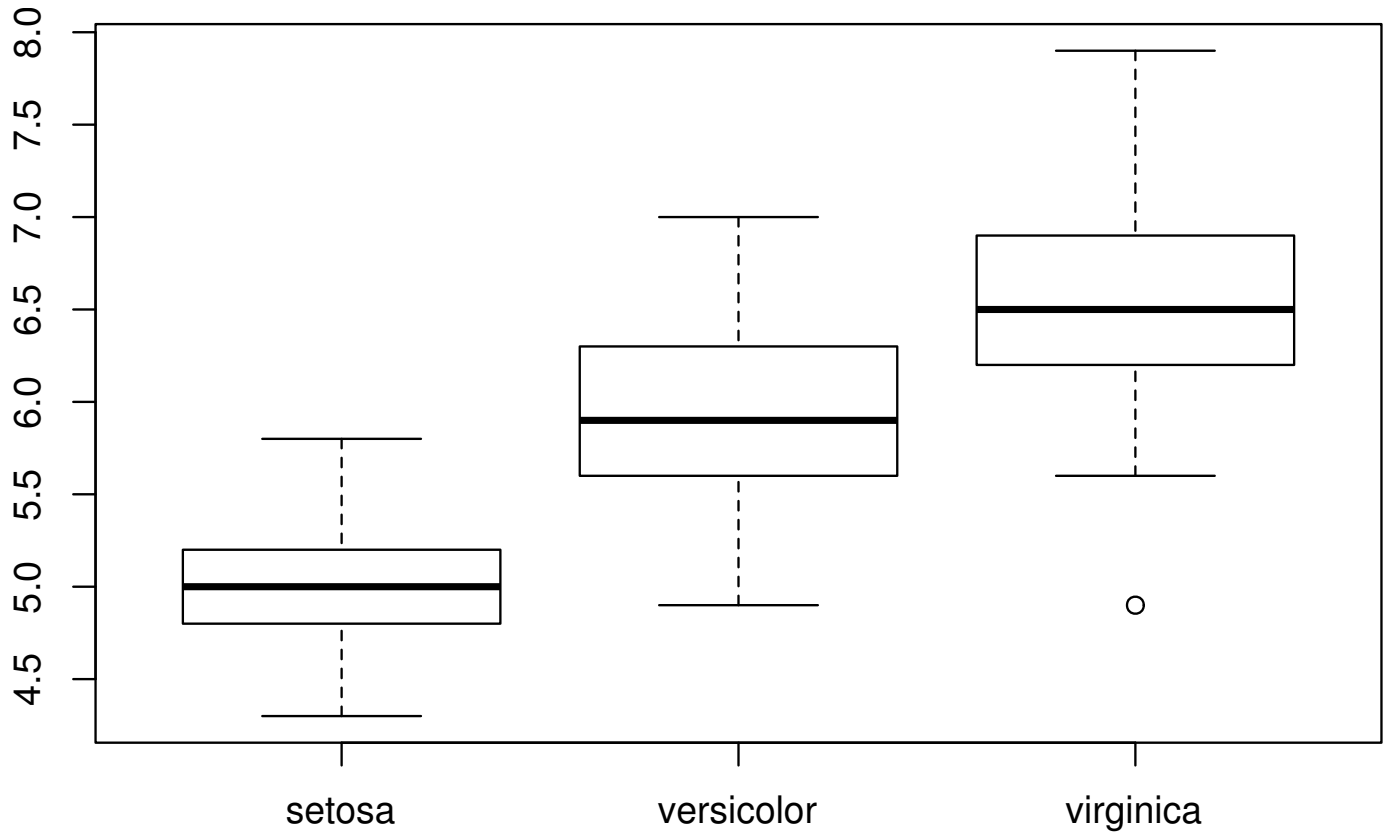
$$Q_3 = \frac{X_{([3\times n/4])} + X_{([3\times n/4]+1)}}{2}$$

3. IQR(interquartile range; 4분위수범위)의 계산: $IQR = Q_3 - Q_1$
4. Box 그리기 (상자의 위 아래의 좌표계산):
   $bottom = Q_1 - 1.5 * IQR, top = Q_3 + 1.5 * IQR$
5. Box밖 수염 (whisker) 그리기: 상자의 아래 혹은 위로 다음의 값까지를 선으로 연결함
   $$w_1 = Q_1 - 1.5 * IQR, \quad w_2 = Q_3 + 1.5 * IQR$$
6. 이상치 판정: $X < Q_1 - 1.5 * IQR$ or $X > Q_3 + 1.5 * IQR$
7. “Extreme" outliers: $X < Q_1 - 3 * IQR$ or $X > Q_3 + 3 * IQR$
8. “Mild” outliers

$$Q_1 - 3 * IQR < X < Q_1 - 1.5 * IQR \text{ or } Q_3 + 1.5 * IQR < X < Q_3 + 3 * IQR$$

예제:

```
boxplot(Sepal.Length~Species, data=iris)
```



확률 및 확률 분포 (probability and probability distribution)

- 확률은 모집단에서 표본을 추출할 때, 표본을 바탕으로 모집단에 대한 결론을 이끌어내는 데 논리적 근거임
  - 표본공간(sample space, S): 통계적 조사 혹은 실험에서 얻을 수 있는 모든 가능한 결과들의 전체집합
  - 사건(event, A): 표본공간의 부분집합
- 확률 (probability) : 어떤 사건이 일어날 가능성을 0과 1사이의 수로 대응시킨 관계로 다음의 성질을 만족함
  1. 임의의 사건 A에 대하여, $0 \leq P(A) \leq 1$
  2. $P(\emptyset) = 0, P(S) = 1$
  3. 서로 배반인 사건열 $\{A_i\}$ 에 대하여 $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.
- 확률변수(random variable): 표본공간 S 에서 정의된 실함수, $X : S \to R$.
- 확률분포(probability distribution): 표본공간에서 정의된 확률 P를 확률변수 X의 값에 따라 재표현한 것으로 X의 확률분포라고 함

$$\text{For all } A \subset S, X(A) = \{X(\omega) : \omega \in A\} \equiv B,$$

$$\text{then } P(A) = \Pr(B) = P(X^{-1}(A)) = P(X \in B).$$

확률분포의 표현

확률(분포)은 원래 집합함수로 정의됨, 이를 쉽게 표현하는 방법

X 의 누적분포함수 (cumulative distribution function:c.d.f.):

$$F : R \to [0, 1], \text{that is,}$$

$$F(x) = P(X \in (-\infty, x]) \equiv P(X \leq x).$$

확률질량함수, 확률밀도함수

- 확률 분포함수가 $(-\infty, x]$ 구간에서의 확률인 반면 확률질량함수(probability mass function:p.m.f.)는 하나의 점 (point)에서의 확률를 표현한 함수

$$p(x) = P(X = x)$$

- 확률 분포함수가 미분 가능이면, 그 도함수를 확률밀도함수(probability density function: pdf)

$$f(x) = \frac{dF(x)}{dx}$$

기대값 (Expectation)

- 확률변수 x의 기대값 ($f(x)$는 확률밀도함수)

$$E(X^r) = \int x^r f(x) dx$$

- 주어진(관측된) 자료를 이용한 기대값의 근사

$$E(X^r) \simeq \frac{1}{n} \sum_{i=1}^{n} x_i^r.$$

  - 표본 평균: $r = 1$인 경우

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

  - 표본분산 : $r = 2$인 경우

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n}{n-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^2 \right\} - \frac{n}{n-1} \bar{X}^2.$$

확률 및 확률 분포 (probability and probability distribution)

- 이산확률분포 (discrete distribution)
  - 베르누이 (Bernoulli)
  - 이항분포 (Binomial), 다항분포 (Multinomial)
  - 포아송 (Poisson)
  - ...
- 연속확률분포 (continuous distribution)
  - 정규분포 (Normal, Gaussian)
  - 지수분포 (Exponential)
  - 감마분포 (Gamma)
  - 베타분포 (Beta)
  - ...

통계적 추론

- 통계적 추론: 연구자가 관심을 가지는 모집단의 특성치 (모수, parameter)에 대한 판단을 하기 위하여, 수집된 데이터를 기초로 통계 이론을 이용한 일련의 과정
  - 추정 (estimation) :
    1. 점추정 : 관측된 표본을 이용하여 모수 (θ)를 하나의 값으로 추정
    2. 구간추정: 관측된 표본을 이용하여 모수가 포함되리라고 예상되는 범위를 추정
    3. 신뢰수준 (confidence level): $1-\alpha$ 또는 $(1-\alpha)\times 100\%$

$$P(L(X) \leq \theta \leq U(X)) = 1-\alpha$$

  - 가설검정 (hypothesis test) : 모수 또는 모집단 분포에 대한 가정 (가설)을 세우고, 표본을 기초로 가정의 참 거짓을 판단하는 방법
    1. 대립가설(alternative hypothesis): 통상적으로 연구자가 입증하려는 가설로, 표본을 토대로 확실한 근거가 있어야 받아들임
    2. 귀무가설(null hypothesis): 대립가설과 상반되는 가설
    3. 보통 귀무가설을 H_0, 대립가설을 H_1으로 표시함
    4. 검정통계량 (test statistic): 가설검정을 하기 위해 이용하는 통계량
    5. 유의수준 (significance level):
    6. 기각역 (critical region): 귀무가설을 기각시키기 위한 검정통계량 관측값의 영역
    7. P값 (유의확률; P-value): 귀무가설이 참이라는 전제하에서 검정통계량이 관측값을 벗어날 확률

$$P(T > t | H_0 \ true)$$

일표본 (단변량)에서 모평균 추정

- 랜덤표본 : $X_1, X_2, \ldots, X_n \sim$ *iid* $N(\mu, \sigma^2)$,
- 모평균 (μ)의 추정

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

- 모분산 (σ^2)의 추정

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

- 모평균 (μ)의 구간추정 (신뢰구간, confidence Interval)
  - 신뢰수준 (confidence level) $1 - \alpha$
  - 모분산이 알려진 경우 (σ^2), 또는 표본의 수가 많은 경우 ($\hat{\sigma}^2$)

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

  - 모분산을 모르는 경우 ($\hat{\sigma}^2$)

$$(\bar{X} - t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}})$$

  - $z_{\alpha/2}$: 표준정규분포에서 $\alpha/2$-백분위수
  - $t_{\alpha/2}$: 자유도가 $n-1$인 t분포의 $\alpha/2$-백분위수

예제: 모평균 추정

- 수면제의 효과 측정 데이터 (sleep) : 10명의 환자에게 수면제를 투여 후, 초과 수면시간 측정
- Objectives: 모평균의 점/구간 추정

```
x <- sleep[1:10, c(3,1)]
head(x)
```

```
##    ID extra
## 1   1   0.7
## 2   2  -1.6
## 3   3  -0.2
## 4   4  -1.2
## 5   5  -0.1
## 6   6   3.4
```

예제: 모평균 추정 (conti)

```
library(Rmisc)
c1 <- CI(x[,2], ci=0.95)[c(3,1)]
c2 <- CI(x[,2], ci=0.99)[c(3,1)]
o <- rbind(c1, c2); row.names(o)<-NULL
o <- data.frame("level"=c(0.95, 0.99), o, mean=c(mean(x[,2]), NA), se=c(sd(x[,2
o
```

```
##   level      lower    upper mean        se
## 1  0.95 -0.5297804 2.029780 0.75 0.5657345
## 2  0.99 -1.0885442 2.588544   NA        NA
```

일표본 (단변량)에서 모비율 추정

- 랜덤표본 : $X_1, X_2, \ldots, X_n \sim$ *iid Bernoulli*(p),
- $\sum_{i=1}^{n} X_i = X_1 + \ldots + X_n$ 의 분포는 (n, p)를 모수로 하는 이항분포를 따름 $(Bin(n, p))$
- 표본비율: 모비율 (p)의 추정

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

- 표본비율의 정규근사 (중심극한정리): $np \geq 5, n(1-p) \geq 5$

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$$

- 모비율의 구간추정 (신뢰수준: $1 - \alpha$)

$$\hat{p} \pm z_{\alpha/2} \frac{\hat{p}(1-\hat{p})}{\sqrt{n}}$$

예제: 모비율 추정

- 대학생의 흡연율을 파악하기 위해 100명을 랜덤 추출, 흡연여부를 조사하여 아래의 결과를 얻었다.

흡연	비흡연	합계
30	70	100

- Objectives: 모비율의 점추정치 및 95% 신뢰구간

예제: 모비율 추정 (conti)

```
x <- 30; n <- 100; p <- x/n;
se <- sqrt(p*(1-p)/n)

z_alpha <- qnorm(0.975)
lb <- p - z_alpha * se
ub <- p + z_alpha * se

o <- t(c(p, se, lb, ub))
colnames(o)<- c("mean", "se", "lower", "upper")
o
```

```
##      mean          se     lower     upper
## [1,]  0.3 0.04582576 0.2101832 0.3898168
```

가설검정에서 오류

	H_0 is TRUE	H_0 is FALSE
Reject H_0	Type I Error	True Positive
Accept H_0	True Negative	Type II error

- Type I Error: 귀무가설이 참인데도 불구하고 귀무가설을 기각할 오류
- 유의수준 (significance level): Type I Error를 범할 확률의 최대값

$$\alpha = P(Reject\ H_0 | H_0\ is\ TRUE)$$

- Type II Error: 귀무가설이 거짓인데도 불구하고 귀무가설을 채택하는 오류
- 검정력 (Power) : 귀무가설이 거짓일 때, 귀무가설을 기각할 확률

$$1 - \beta = P(Reject\ H_0 | H_0\ is\ FALSE)$$

t검정

1. 일표본 t검정
2. 이표본 t검정
3. 대응비교

일표본 t검정

- 표본자료: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- 모평균이 μ_0 (특정값)인지를 관측된 표본을 이용하여 검증
- 가설: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- 검정통계량 :
  1. 모집단 분산(σ^2)이 알려진 경우

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ :표준정규분포}$$

  2. 모집단 분산을 모르는 경우: 자유도(degree of freedom)가 $n-1$인 t분포

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1),$$

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

- 기각역: 유의수준 α인 양측검정에서

$$|T| > t_{\alpha/2, n-1}$$

- p-value(유의 확률): 검정통계량의 관측값이 t_0일 때,

$$p - value = \Pr(|T| > t_0)$$

예제: 일표본 t검정

- 수면제의 효과 측정 데이터 (sleep) : 10명의 환자에게 수면제를 투여 후, 초과 수면시간 측정
- 가설 : 초과 수면시간(μ)이 0보다 큰가? ($H_0 : \mu = 0$ vs, $H_1 : \mu > 0$)

```
x <- sleep[1:10, c(3,1)]
head(x)
```

```
##   ID extra
## 1  1   0.7
## 2  2  -1.6
## 3  3  -0.2
## 4  4  -1.2
## 5  5  -0.1
## 6  6   3.4
```

예제: 일표본 t검정 (conti.)

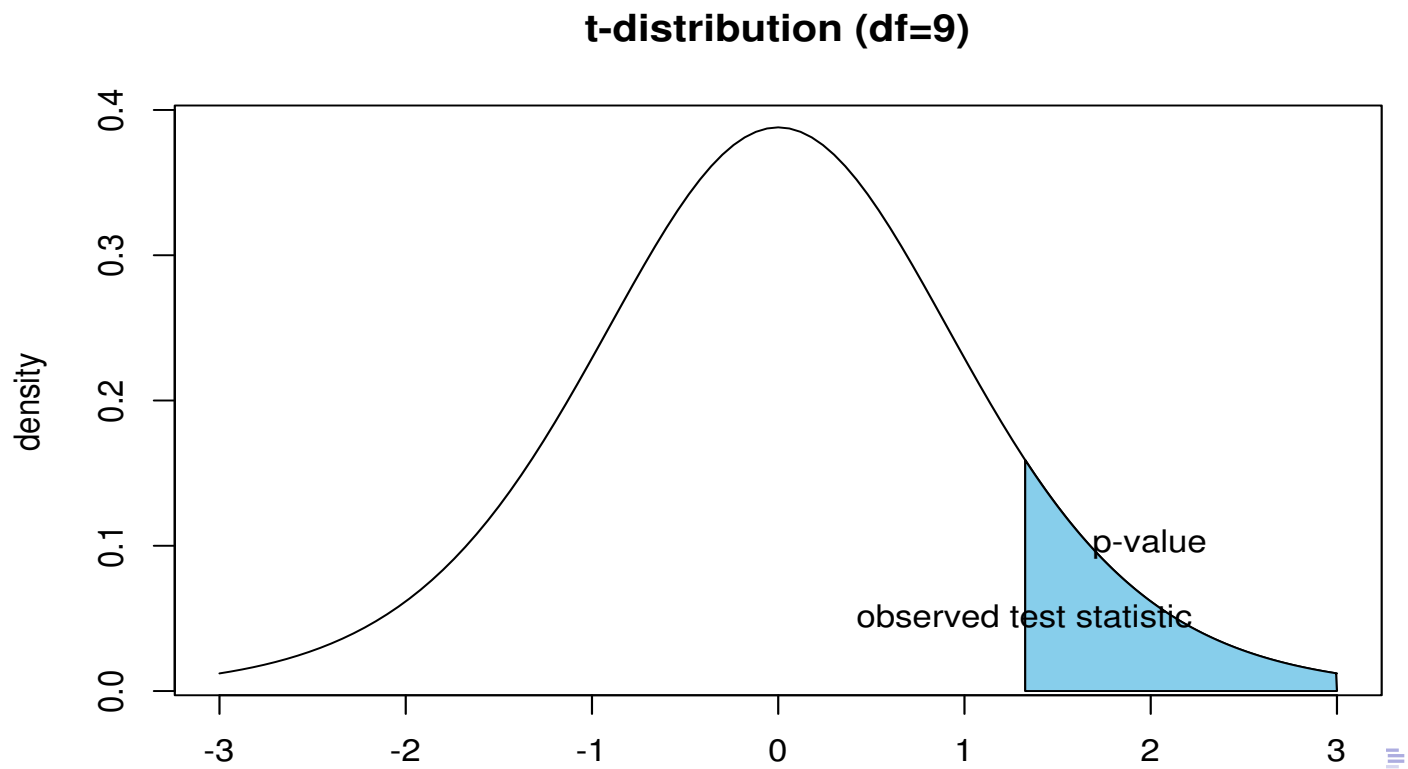
- alternative="greater": $H_1 : \mu > 0$

```
o <- t.test(x=x$extra, alternative="greater")
o
```

```
##
##  One Sample t-test
##
## data:  x$extra
## t = 1.3257, df = 9, p-value = 0.1088
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.2870553        Inf
## sample estimates:
## mean of x
##      0.75
```

예제: 일표본 t검정 (conti.)

```
cord.x <- c(o$statistic, seq(o$statistic, 3, 0.01), 3)
cord.y <- c(0, dt(seq(o$statistic, 3, 0.01), 9), 0)
curve(dt(x, 9), xlim=c(-3,3), ylab="density", main='t-distribution (df=9)')
polygon(cord.x,cord.y,col='skyblue')
text(o$statistic, 0.05, "observed test statistic")
text(2, 0.1, "p-value")
```



이표본(독립표본) t검정 (Two Sample t-test) : 두 모집단의 평균의 차이 검정

- 표본자료

$$\begin{aligned} X_1, \ldots, X_m &\sim N(\mu_1, \sigma_1^2) \\ Y_1, \ldots, Y_n &\sim N(\mu_2, \sigma_2^2) \end{aligned}$$

- 가설 : $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$
- 검정통계량:

$$T = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}}, \text{ where } s_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}$$

- 검정통계량의 분포
  1. 분산이 서로 같은 경우 : $t(n+m-2)$
  2. 다른 경우: Welch t 검정이며 자유도는

$$df = \frac{(s_x^2/n + s_y^2/m)^2}{(s_x^2/m)^2/(m-1) + (s_y^2/n)^2/(n-1)}$$

예제: 독립표본 t검정

- 수면제의 효과 측정 데이터 (sleep) : 수면제 A, B를 각 10명의 환자에게 투여 후, 초과 수면시간 측정
- 가설 : A의 초과 수면시간(μ_1) 보다 B(μ_2)가 큰가? ($H_0 : \mu_1 = \mu_2$ vs, $H_1 : \mu_1 < \mu_2$)

```
A <- sleep[1:10, 1]
B <- sleep[11:20, 1]
x <- data.frame(n=1:10, A, B)
x
```

```
##      n    A    B
## 1    1  0.7  1.9
## 2    2 -1.6  0.8
## 3    3 -0.2  1.1
## 4    4 -1.2  0.1
## 5    5 -0.1 -0.1
## 6    6  3.4  4.4
## 7    7  3.7  5.5
## 8    8  0.8  1.6
## 9    9  0.0  4.6
## 10 10  2.0  3.4
```

예제: 독립표본 t검정 (conti.)

- 이분산: var.equal=FALSE

```
t.test(x=x$A, y=x$B, paired=F, alternative="less", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x$A and x$B
## t = -1.8608, df = 17.776, p-value = 0.0397
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1066185
## sample estimates:
## mean of x mean of y
##      0.75      2.33
```

예제: 독립표본 t검정 (conti.)

- 등분산: `var.equal=TRUE`

```
t.test(x=x$A, y=x$B, paired=F, alternative="less", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x$A and x$B
## t = -1.8608, df = 18, p-value = 0.03959
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1076222
## sample estimates:
## mean of x mean of y
##      0.75      2.33
```

예제: 독립표본 t검정 (conti.)

- 등분산 검정
- 가설 : $H_0 : \sigma_1^2/sigma_2^2 = 1$ vs. H_1: not H_0
- 검정통계량 : $F = s_x^2/s_y^2 \sim F(m-1, n-1)$

```
var.test(A, B)
```

```
##
##  F test to compare two variances
##
## data:  A and B
## F = 0.79834, num df = 9, denom df = 9, p-value =
## 0.7427
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.198297 3.214123
## sample estimates:
## ratio of variances
##          0.7983426
```

- 분산이 다르지 않음 : 등분산 가정의 t검정 결과를 이용함

대응비교(짝비교, Paired t-test)

- 동일한 대상에 대하여 서로 다른 처리를 한 후 처리 효과의 차이를 비교할 때
- 표본자료:

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim N\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \Sigma\right)$$

Then

$$D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n \sim N(\mu_1 - \mu_2, \sigma_D^2)$$

- 검정통계량:

$$T = \frac{\bar{D}}{s_D/\sqrt{n}} \sim t(n-1), s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

예제: 대응비교 t검정

- 운동선수 트레이닝 방법 : 100m 단거리 육상선수 10명에게 새로운 훈련방법을 도입하여 전/후 효과를 측정
- 가설 : 훈련 후 (μ_2)가 훈련 전 (μ_1)보다 효과가 있는가? ($H_0 : \mu_1 = \mu_2$ vs, $H_1 : \mu_1 < \mu_2$)

```
pre = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
post = c(12.7, 13.6, 12.0, 15.2, 16.8, 20.0, 12.0, 15.9, 16.0, 11.1)
x <- data.frame(id=1:10, pre, post, d=post-pre)
x
```

```
##    id  pre post    d
## 1   1 12.9 12.7 -0.2
## 2   2 13.5 13.6  0.1
## 3   3 12.8 12.0 -0.8
## 4   4 15.6 15.2 -0.4
## 5   5 17.2 16.8 -0.4
## 6   6 19.2 20.0  0.8
## 7   7 12.6 12.0 -0.6
## 8   8 15.3 15.9  0.6
## 9   9 14.4 16.0  1.6
## 10 10 11.3 11.1 -0.2
```

예제: 대응비교 t검정

```r
t.test(post-pre, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  post - pre
## t = 0.21331, df = 9, p-value = 0.4179
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.3796859        Inf
## sample estimates:
## mean of x
##      0.05
```

비율검정

- one sample 비율검정

$$X_1, \ldots, X_n \sim \text{iid } Ber(p) \Rightarrow X = \sum X_i \sim Bin(n, p)$$

- Hyperthesis:

$$H_0 : p = p_0, \text{vs } H_1 : \text{ not } H_0$$

- Test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1) \text{ under } H_0$$

비율검정 (conti)

- 예제 : 대학생의 흡연율이 40%보다 작은가?
  - 가설: $H_0 : p = p_0$, vs H_1 : not H_0
  - 결과

```
x <- 30; n <- 100; hatp <- x/n; p0 <-0.4
z <- (hatp - p0)/sqrt(p0*(1-p0)/n)
prop.test(x, n=n, p=p0, correct=F, alternative="l")
```

```
##
##  1-sample proportions test without continuity
##  correction
##
## data:  x out of n, null probability p0
## X-squared = 4.1667, df = 1, p-value = 0.02061
## alternative hypothesis: true p is less than 0.4
## 95 percent confidence interval:
##  0.0000000 0.3798321
## sample estimates:
##   p
## 0.3
```

비율검정 (이표본 비교)

- Data

	group 1	group 2
success	X_1	X_2
trials	n_1	n_2
success probability	$\hat{p}_1 = X_1/n_1$	$\hat{p}_2 = X_2/n_2$

- Hyperthesis:

$$H_0 : p_1 = p_2, \text{vs } H_1 : \text{ not } H_0$$

- Test statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}.$$

예제 : 비율검정 (이표본 비교)

- 특정 바이러스에 대한 남, 여 항체보유율을 비교,

	남	여	합계
대상자	100	150	250
항체보유자	24	64	58

- 가설: $H_0 : p_1 = p_2$ vs $H_1 : p_1 < p_2$

예제 : 비율검정 (이표본 비교) - conti

```
x <- c(24, 64); n <- c(100, 150);
p <- x/n;
prop.test(x, n=n, correct=F, alternative="l")
```

```
##
##  2-sample test for equality of proportions without
##  continuity correction
##
## data:  x out of n
## X-squared = 9.1657, df = 1, p-value = 0.001233
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.000000 -0.089986
## sample estimates:
##    prop 1    prop 2
## 0.2400000 0.4266667
```

- 여자의 항체보유율이 남자에 비해 높다고 판단