07 강

R 데이터 분석

통계추론 II

한림대학교 데이터사이언스학부 심송용 교수

한국방송통신대학교
Korea National Open University

학습목차

1. 분산에 대한 추론
2. 상관계수에 대한 추론
3. 적합도와 독립성 검정
4. 단순회귀분석
5. 일원배치 분산분석

3.6.1. 일표본 분산 추론

$X_1, X_2, \ldots, X_n$이 평균 μ, 분산 σ^2인 정규분포 확률표본
표본평균 및 분산을 각각 $\overline{X}$, S^2 이라고 하면

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

은 자유도 $(n-1)$인 카이제곱분포. 따라서 모분산에 대한
$100(1-\alpha)\%$ 신뢰구간은

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \right)$$

로 얻으며 귀무가설 $H_0 : \sigma^2 = \sigma_0^2$ 에 대한 가설검정은

3.6.1. 일표본 분산 추론

검정통계량

$$\chi_0^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma_0^2}$$

을 사용. 각 대립가설에 따른 유의확률 및 기각역

대립가설	기각역	유의확률 P
$H_1 : \sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi_{n-1;\alpha}^2$	$\Pr[X_{n-1}^2 > \chi_0^2]$
$H_1 : \sigma^2 < \sigma_0^2$	$\chi_0^2 < \chi_{n-1;1-\alpha}^2$	$\Pr[X_{n-1}^2 < \chi_0^2]$
$H_1 : \sigma^2 \neq \sigma_0^2$	$\chi_0^2 > \chi_{n-1;\alpha/2}^2$ 또는 $\chi_0^2 < \chi_{n-1;1-\alpha/2}^2$	$\chi_0^2 < 1$ 이면 $2\Pr[\chi_{n-1}^2 < \chi_0^2]$ $\chi_0^2 > 1$ 이면 $2\Pr[\chi_{n-1}^2 > \chi_0^2]$

3.6.1. 일표본 분산 추론

예제 3.14: R에서 분산이 4인 정규분포 난수 100개를 발생하여 분산이 4보다 크지 유의수준 5%에서 검정하고 95% 신뢰구간 계산

(var1s.test.r)

```
> nn <- 100
> x <- rnorm(nn, sd =2); vx <- var(x)
> chi0 <- (nn-1)*vx / 2^2  ➡ 103.5662
> p.val <- 1-pchisq(chi0, nn-1)  ➡ 0.3568574
> ci <- c( (nn-1)*vx /qchisq(0.975, nn-1),
           (nn-1)*vx /qchisq(0.025, nn-1) )
> ci
[1] 3.225810 5.646932
```

3.6.2. 이표본 분산 추론

$X_1, X_2, \ldots, X_m$: $N(\mu_1, \sigma_1^2)$ 의 확률표본
$Y_1, Y_2, \ldots, Y_n$: $N(\mu_2, \sigma_2^2)$ 의 확률표본
두 확률표본도 독립

S_1^2 과 S_2^2 이 각각 X 와 Y 들의 표본분산이면

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

는 자유도가 $(m-1, n-1)$ 인 F-분포. 따라서 $\dfrac{\sigma_1^2}{\sigma_2^2}$ 의 $100(1-\alpha)\%$ 신뢰구간은

$$\left(\frac{1}{F_{m-1,n-1;\alpha/2}} \frac{s_1^2}{s_2^2}, \frac{1}{F_{m-1,n-1;1-\alpha/2}} \frac{s_1^2}{s_2^2} \right)$$

3.6.2. 이표본 분산 추론

이고 귀무가설 $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = r$ 에 대한 검정통계량은

$$F_0 = \frac{S_1^2}{rS_2^2}$$

이고 귀무가설이 참이면 자유도 (m-1, n-1)인 F-분포.

기각역 및 유의확률

대립가설	기각역	유의확률 P
$H_1 : \sigma_1^2/\sigma_2^2 > r$	$F_0 > F_{m-1,n-1;\alpha}$	$\Pr[F_{m-1,n-1} > F_0]$
$H_1 : \sigma_1^2/\sigma_2^2 < r$	$F_0 < F_{m-1,n-1;1-\alpha}$	$\Pr[F_{m-1,n-1} < F_0]$
$H_1 : \sigma_1^2/\sigma_2^2 \neq r$	$F_0 > F_{m-1,n-1;\alpha/2}$ 또는 $F_0 < F_{m-1,n-1;1-\alpha/2}$	$F_0 < 1$ 이면 $2\Pr[F_{m-1,n-1} < F_0]$ $F_0 > 1$ 이면 $2\Pr[F_{m-1,n-1} > F_0]$

3.6.3. 분산추론의 R-함수

```
var.test(x, y, ratio = 1, alternative = c("two.sided", "less",
            "greater"), conf.level = 0.95, ...)
또는
var.test(formula, data, subset, na.action,
```

예제 3.16: 예제 3.2의 자료를 var.test 함수에 적용하면 다음과 같은 결과를 얻음.

```
> 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)
> var.test(x,y)
```

3.6.3. 분산추론의 R-함수

```
F test to compare two variances

data:  x and y
F = 3.8723, num df = 9, denom df = 10,
                                    p-value = 0.04617
alternative hypothesis: true ratio of variances is not
                                   equal to 1
95 percent confidence interval:
  1.024708 15.349414
sample estimates:
ratio of variances
         3.872335
```

3.7. 상관계수에 대한 추론

$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ 이 독립인 이변량 정규분포의 확률벡터.
두 확률변수 X와 Y의 표본 공분산

$$Cov(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

로 얻으며, S_x^2과 S_y^2을 각각 X와 Y의 표본분산이라고 하면, 표본 피어슨 선형상관계수:

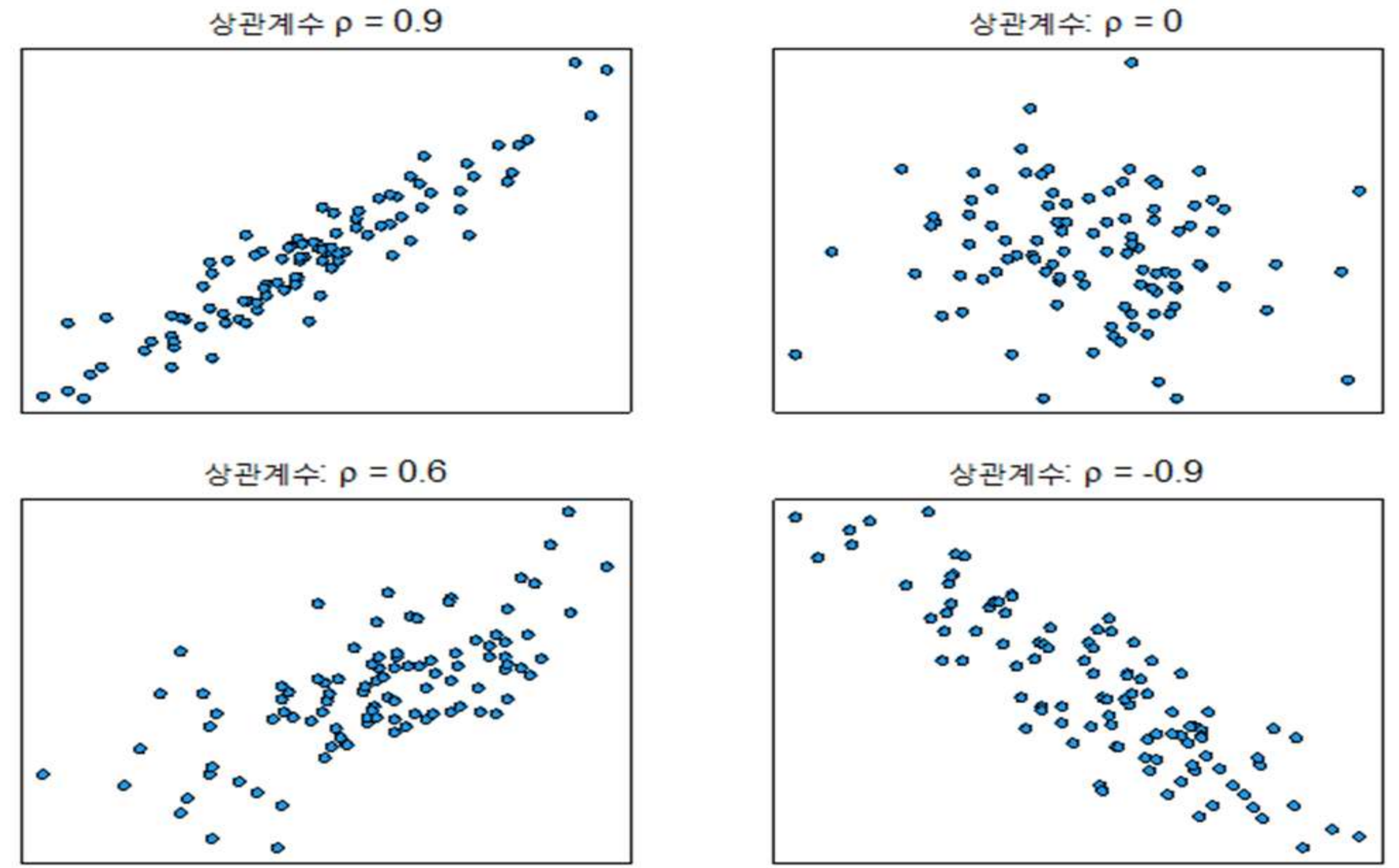
$$Corr(X, Y) = \frac{Cov(X, Y)}{S_x S_y} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}$$

3.7. 상관계수에 대한 추론

상관계수 r 의 성질

- $-1 \leq r \leq 1$ 의 값.
- x가 증가(감소)할 때 y가 증가(감소)하면 상관계수는 양의 값을 가지며, x가 증가(감소)할 때 y가 감소(증가)하면 상관계수는 음의 값.
- 상관이 높을수록 상관계수의 절댓값이 1에 가까워지며, 자료의 모든 점이 한 직선 위에 존재하면 상관계수는 1 또는 -1.
- 두 변수가 독립이면 표본에서 얻은 상관계수 은 0에 가까운 값을 가지며, 역은 성립하지 않음. 즉 상관계수가 0(또는 0에 가까운 값)이라고 하더라도 두 변수는 독립이 아닐 수 있음.

3.7. 상관계수에 대한 추론



3.7.1. Fisher 변환에 의한 상관계수에 대한 추론

Fisher 변환: 모상관계수가 ρ 이고 표본상관계수가 r 일 때

$$\nu = \frac{1}{2}\log\frac{1+r}{1-r} \sim N\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$$

따라서 귀무가설 $H_0 : \rho = \rho_0$ 에 대한 가설검정의 검정통계량은

$$Z = \frac{\frac{1}{2}\log\frac{1+r}{1-r} - \frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}}{\sqrt{1/(n-3)}}$$

3.7.1. Fisher 변환에 의한 상관계수에 대한 추론

(근사) 기각역 및 유의확률

대립가설	기각역	유의확률 P
$H_1 : \rho > \rho_0$	$z_0 > z_\alpha$	$\Pr[Z > z_0]$
$H_1 : \rho < \rho_0$	$z_0 < -z_\alpha$	$\Pr[Z < z_0]$
$H_1 : \rho \neq \rho_0$	$\lvert z_0 \rvert > z_{\alpha/2}$	$2\Pr[Z > \lvert z_0 \rvert]$

신뢰구간은 근사정규분포인 $\nu = \frac{1}{2}\log\frac{1+r}{1-r}$ 의 신뢰구간을 구한 후 역변환 $r = \frac{e^{2\nu}-1}{e^{2\nu}+1}$ 으로 얻음

3.7.2. 회귀계수의 추론에서 유도된 추론

단순선형 회귀분석의 기울기에 대한 추론에서 상관계수에 대한 추론을 얻음.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

은 $H_0 : \rho = 0$가 참일 때 자유도가 $(n-2)$인 t-분포.

귀무가설 에 대한 기각역 및 유의확률

대립가설	기각역	유의확률 P
$H_1 : \rho > 0$	$t_0 > t_{n-2;\alpha}$	$\Pr[T_{n-2} > t_0]$
$H_1 : \rho < 0$	$t_0 < -t_{n-2;\alpha}$	$\Pr[T_{n-2} < t_0]$
$H_1 : \rho \neq 0$	$\lvert t_0 \rvert > t_{n-2;\alpha/2}$	$2\Pr[T_{n-2} > \lvert t_0 \rvert]$

3.7.3. 상관계수 추론을 위한 R 함수

```
cor.test(x, y, alternative = c("two.sided", "less",
       "greater"), method = c("pearson", "kendall",
       "spearman"), conf.level = 0.95,  ...)
```

예제 3.18: 아이의 재능에 대한 부모의 평가와 교사의 평가를 조사하였더니 다음과 같았다. 상관계수가 0인지 검정하고, 상관계수에 대한 95% 신뢰구간을 얻어 보자.

부모평가	35	35	33	34	31	35	35	35	35	35	33	35	35	35	31	32	35
교사평가	25	31	33	33	34	33	34	33	29	33	35	35	35	35	35	35	32

3.7.3. 상관계수 추론을 위한 R 함수

```
> x <- c(35,35,33,34,31,35,35,35,35,35,33,35,35,…)
> y <- c(25,31,33,33,34,33,34,33,29,33,35,35,35,…)
> cor.test(x,y)
        Pearson's product-moment correlation

data:  x and y
t = -1.5351, df = 15, p-value = 0.1456
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7213589  0.1363185
sample estimates:
       cor
-0.3684686
```

3.8.1. 적합도 검정

k개의 범주에 대해서 i번째 범주에 속할 확률이 p_i인지 검정하는 문제.
자료를 요약하면 다음과 같은 표.

범주	1	2	$\cdots$	k	합
빈도	O_1	O_2	$\cdots$	O_k	n

i번째 범주에 속할 확률이 p_i라면 i번째 범주에 대한 기대 도수는 $E_i = np_i$. 귀무가설 'H_0 : i번째 범주의 확률은 p_i 이다' 대 대립가설 'H_0가 아니다'에 대한 검정통계량

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

은 귀무가설이 참일 때 근사적으로 자유도가 $(k-1)$인 카이제곱분포를 따름(모든 기대도수 E_i가 5 이상인 경우)

3.8.2. 독립성 검정

- 각각 r 개와 c개의 범주를 가진 두 범주형 자료의 독립성 검정
- 두 변수는 편의상 행변수 및 열변수로
- 행변수의 i번째 범주 및 열변수의 j번째 범주의 관측빈도수: O_{ij}.
- $O_{i.} = \sum_{j=1}^{c} O_{ij}$, $O_{.j} = \sum_{i=1}^{r} O_{ij}$: 각각 i번째 행, j번째 열의 빈도합
- 전체 자료수는 $O_{..} = \sum_{i=1}^{r} O_{i.} = \sum_{j=1}^{c} O_{.j} = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}$
- 자료를 요약하면 교차표(분할표)

3.8.2. 독립성 검정

구분	1	2	$\cdots$	c	합
1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1.}$
2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r.}$
합	$O_{.1}$	$O_{.2}$	$\cdots$	$O_{.c}$	$O_{..}$

행변수와 열변수가 독립일 때 번째 (i,j)칸의 기대도수 $E_{ij} = \frac{O_{i.}O_{.j}}{O_{..}}$

귀무가설 H_0: '행변수와 열변수가 독립이다' 가 참이면

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

는 근사적으로 자유도 (r-1)(c-1)인 카이제곱분포.

3.8.3. 적합도 및 독립성 검정을 위한 R 함수

```
chisq.test(x, y = NULL, correct = TRUE,
              p = rep(1/length(x), length(x)), …)
```

예제 3.21: 주사위를 100번 던져 기록한 눈금의 횟수가 다음과 같을 때 이 주사위의 각 눈금이 나올 확률이 모두 1/6인지 검정.

```
> frq <- c(19, 16,19,18, 14, 14)
> chisq.test(frq)

        Chi-squared test for given probabilities

data:  frq
X-squared = 1.64, df = 5, p-value = 0.8964
```

3.8.3. 적합도 및 독립성 검정을 위한 R 함수

성별과 정당지지도를 조사한 다음 결과에서 성별과 정당지지도가 독립이라고 할 수 있겠는가?

	정당 A	정당 B	정당 C	합
남	20	30	15	65
여	30	20	15	65
합	50	50	30	130

```
> obs <- matrix(c(20,30,15,30,20,15), ncol=3, byrow=T)
> chisq.test(obs)

        Pearson's Chi-squared test

data:  obs
X-squared = 4, df = 2, p-value = 0.1353
```

3.9.1. 회귀분석 개요

주어진 값 x_i 에서 Y_i 가

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \ldots, n$$

인 관계에서 회귀계수인 절편 β_0 와 기울기 β_1 을 추정하고 관련된 추론을 하는 문제

오차항 $\epsilon_i = Y_i - \beta_0 - \beta_1 x_i$ 의 제곱합을 최소로 하는 방법으로 추정을 주로 사용 (최소제곱법(Least Squares Method))

$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, \ S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})$ 이라 할 때 최소제곱법에

의한 회귀계수 추정량은

$$b_1 = \widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}, \qquad b_0 = \widehat{\beta_0} = \overline{Y} - b_1 \overline{x}$$

3.9.1. 회귀분석 개요

x_i 에서 종속변수 Y 의 예측치: $\hat{Y}_i = b_0 + b_1 x_i$

제곱합의 분해

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \quad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

요인	제곱합	자유도	평균제곱	F	유의확률
회귀	SSR	1	MSR=SSR/1	$F_0 = \frac{MSR}{MSE}$	$\Pr[F_{1;n-2} > F_0]$
잔차	SSE	$n-2$	MSE = SSE/$(n-2)$		
전체	SST	$n-1$			

$F_0 > F_{1,n-2;\alpha}$ 이거나 유의확률이 유의수준보다 작으면 귀무가설 $H_0 : \beta_1 = 0$ 을 기각하고 대립가설 $H_1 : \beta_1 \neq 0$ 을 채택

3.9.1. 회귀분석 개요

오차항의 분산 σ^2의 추정량은 분산분석표의 MSE

$$\hat{\sigma}^2 = \mathrm{MSE} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

추정량 b_1 과 b_0의 분포

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \quad \text{및} \quad b_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]\right)$$

σ^2 대신 MSE 사용하면

$$\frac{b_1 - \beta_1}{\sqrt{\mathrm{MSE}/S_{xx}}} \sim t_{n-2} \quad \text{및} \quad \frac{b_0 - \beta_0}{\sqrt{\mathrm{MSE}\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}} \sim t_{n-2}$$

3.9.1. 회귀분석 개요

회귀계수의 신뢰구간

$$\beta_1 = b_1 \pm t_{n-2;\alpha/2} \sqrt{\frac{\mathrm{MSE}}{S_{xx}}}$$

$$\beta_0 = b_0 \pm t_{n-2;\alpha/2} \sqrt{\mathrm{MSE}\left[\frac{1}{\mathrm{n}} + \frac{\overline{\mathrm{x}}^2}{\mathrm{S_{xx}}}\right]}$$

회귀계수 가설검정: 귀무가설 $H_0 : \beta_1 = \beta_1^0$ 에 대한 검정통계량

$$t_0 = \frac{\mathrm{b}_1 - \beta_1^0}{\sqrt{\mathrm{MSE/S_{xx}}}}$$

기각역 및 유의확률

3.9.1. 회귀분석 개요

대립가설	기각역	유의확률 P
$H_1 : \beta_1 > \beta_1^0$	$t_0 > t_{n-2;\alpha}$	$\Pr[T_{n-2} > t_0]$
$H_1 : \beta_1 < \beta_1^0$	$t_0 < -t_{n-2;\alpha}$	$\Pr[T_{n-2} < t_0]$
$H_1 : \beta_1 \neq \beta_1^0$	$\lvert t_0 \rvert > t_{n-2;\alpha/2}$	$2\Pr[T_{n-2} > \lvert t_0 \rvert]$

$H_0 : \beta_0 = \beta_0^0$ 에 대한 가설검정의 검정통계량

$$t_0 = \frac{b_0 - \beta_0^0}{\sqrt{\mathrm{MSE}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}$$

기각역 및 유의확률

3.9.1. 회귀분석 개요

대립가설	기각역	유의확률 P
$H_1 : \beta_0 > \beta_0^0$	$t_0 > t_{n-2;\alpha}$	$\Pr[T_{n-2} > t_0]$
$H_1 : \beta_0 < \beta_0^0$	$t_0 < -t_{n-2;\alpha}$	$\Pr[T_{n-2} < t_0]$
$H_1 : \beta_0 \neq \beta_0^0$	$\lvert t_0 \rvert > t_{n-2;\alpha/2}$	$2\Pr[T_{n-2} > \lvert t_0 \rvert]$

주어진 x 값에서의 Y의 기댓값의 추정량: $\hat{\mu}_{Y|x} = b_0 + b_1 x$

분포:

$$\hat{\mu}_{Y|x} \sim N\left(\beta_0 + \beta_1 x, \sigma^2\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right]\right)$$

신뢰구간

$$\mu_{Y|x} = \hat{\mu}_{Y|x} \pm t_{n-2;\alpha}\sqrt{\mathrm{MSE}\left[\frac{1}{\mathrm{n}} + \frac{(\mathrm{x}-\overline{\mathrm{x}})^2}{\mathrm{S_{xx}}}\right]})$$

3.9.1. 회귀분석 개요

귀무가설 $H_0 : \mu_{Y|x} = \mu^0_{Y|x}$ 에 대한 검정통계량:

$$t_0 = \frac{\hat{\mu}_{Y|x} - \mu^0_{Y|x}}{\sqrt{\mathrm{MSE}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]}}$$

기각역 및 유의확률

대립가설	기각역	유의확률 P
$H_1 : \mu_{Y|x} > \mu^0_{Y|x}$	$t_0 > t_{n-2;\alpha}$	$\Pr[T_{n-2} > t_0]$
$H_1 : \mu_{Y|x} < \mu^0_{Y|x}$	$t_0 < -t_{n-2;\alpha}$	$\Pr[T_{n-2} < t_0]$
$H_1 : \mu_{Y|x} \neq \mu^0_{Y|x}$	$|t_0| > t_{n-2;\alpha/2}$	$2\Pr[T_{n-2} > |t_0|]$

3.9.1. 회귀분석 개요

주어진 x 값에서의 Y의 예측값: $\widehat{Y_{|x}} = b_0 + b_1 x$

분포:

$$\widehat{Y_{|x}} \sim N\left(\beta_0 + \beta_1 x, \sigma^2\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right]\right)$$

신뢰구간

$$Y = \widehat{Y_{|x}} \pm t_{n-2;\alpha}\sqrt{\mathrm{MSE}\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right]}$$

3.9.2. 회귀분석을 위한 R 함수

```
lsfit(x, y, intercept = TRUE, ... )
```

```
lm(formula, data,  ...)
```

예제 3.23: 다음은 11명의 나이와 혈중 콜레스테롤 농도를 조사한 자료이다. 나이를 독립변수, 콜레스테롤 수치를 종속변수로 회귀분석을 하여 F 검정통계량, 계수의 신뢰구간, 나이에 따른 종속변수의 기댓값에 대한 신뢰구간, 종속변수의 예측구간을 구해 보자.

나이	54	69	43	39	64	52	47	34	73	37	45
콜레스테롤	181	235	193	177	197	191	213	167	212	183	190

3.9.2. 회귀분석을 위한 R 함수

```
> age <- c( 54,  69,  43,  39,  64,  52,  …,  73,  37,  45)
> c.level <- c(181, 235, 193, 177,   …, 212, 183, 190)
> cdata <- data.frame(age, c.level)
> lsfit(age, c.level)
$coefficients
 Intercept          X
138.689641   1.101282
 (출력 일부 생략)

> summary(lm(c.level ~ age, data=cdata))
Call:
lm(formula = c.level ~ age, data = cdata)
```

3.9.2. 회귀분석을 위한 R 함수

```
Coefficients:
            Estimate   Std. Error  t value Pr(>|t|)
(Intercept) 138.6896   16.7613   8.274 1.69e-05 ***
age           1.1013    0.3213   3.428  0.00754 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.37 on 9 degrees of freedom
Multiple R-squared:  0.5662,    Adjusted R-squared:  0.518
F-statistic: 11.75 on 1 and 9 DF,  p-value: 0.00753
```

3.10 일원배치 분산분석에 대한 추론

g 개의 정규분포에서 얻은 자료의 평균이 모두 같은지 검정
i번째 그룹에서 n_i개의 자료를 얻음
전체 자료의 수: $n = n_1 + n_2 + \cdots + n_g$
i번째 그룹의 j번째 자료: Y_{ij}
i번째 그룹의 표본평균: $\overline{Y_{i.}}$, 전체 자료의 평균: $\overline{Y_{..}}$

그룹	자료	평균
1	$Y_{11}, Y_{12}, \cdots, Y_{1n_1}$	$\overline{Y_{1.}}$
2	$Y_{21}, Y_{22}, \cdots, Y_{2n_2}$	$\overline{Y_{2.}}$
$\vdots$	$\vdots$	$\vdots$
g	$Y_{g1}, Y_{g2}, \cdots, Y_{gn_g}$	$\overline{Y_{g.}}$
전체		$\overline{Y_{..}}$

3.10 일원배치 분산분석에 대한 추론

$Y_{ij} \sim N(\mu_i, \sigma^2)$ 이고 모든 자료는 독립. 즉,

1. 정규성: 자료는 모두 정규분포
2. 독립성: 모든 자료는 독립
3. 등분산성: i번째 그룹의 평균은 μ_i 로 다를 수 있으나 분산은 σ^2 으로 모두 같음

귀무가설 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$

대립가설 H_1 : 적어도 한 그룹의 평균은 나머지 그룹과 다름

제곱합의 분해

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{Y}_{i.} - \overline{Y}_{..})^2$$

$$\text{SST} = \text{SSE} + \text{SSTrt}$$

3.10 일원배치 분산분석에 대한 추론

각제곱합의 자유도는
$(n-1)$, $(n-g)$ 및 $(g-1)$

각 제곱합을 해당 자유도로 나눈값을 평균제곱이라 하며
MSE = SSE/(n-g), MSTrt =SSTrt/(g-1)

앞의 가설을 검정하기 위한 검정통계량 $F_0 = \frac{\mathrm{MSTrt}}{\mathrm{MSE}}$ 는 귀무가설이

참이면 자유도 $(g-1, n-g)$ 인 F-분포이므로

$F_0 > F_{g-1,n-g;\alpha}$ 이거나 유의확률이 유의수준보다 작으면 귀무가설 기각

3.10 일원배치 분산분석에 대한 추론

요인	제곱합	자유도	평균제곱	F	유의확률
처리	SSTrt	$g-1$	MSTrt=SSTrt/($g-1$)	$F_0=\dfrac{\text{MSTrt}}{\text{MSE}}$	$\Pr[F_{g-1;n-g} > F_0]$
오차	SSE	$n-g$	MSE = SSE/($n-g$)		
전체	SST	$n-1$			

```
oneway.test(formula, data, subset, na.action,
                    var.equal = FALSE)
```

예제 3.25: 네 가지 비료를 사용하여 얻은 수확량이 다음과 같았다. 비료에 따라 수확량이 차이가 난다고 할 수 있는지 분산분석표를 작성하여 검정해 보자.

3.10 일원배치 분산분석에 대한 추론

비료1	11	11	10	10	10	11	10	8	10	9
비료2	12	11	11	13	12	11	11	11	12	9
비료3	12	13	13	11	10	13	11	12	13	11
비료4	9	10	8	10	13	10	10	10	10	8

```
# oneway.test.r
y <- c( 11, 11, 10, 10, 10, 11, 10,  8, 10,  9,
         12, 11, 11, 13, 12, 11, 11, 11, 12,  9,
         12, 13, 13, 11, 10, 13, 11, 12, 13, 11,
          9, 10,  8, 10, 13, 10, 10, 10, 10,  8)
x <- c(rep(1,10), rep(2,10), rep(3,10), rep(4,10))
fert <- data.frame(y,x)
oneway.test(y ~x, var.equal=T)
```

3.10 일원배치 분산분석에 대한 추론

```
One-way analysis of means

data:  y and x
F = 7.9571, num df = 3, denom df = 36,
p-value = 0.0003382
```

08 강

다음시간 안내

R 통계 그래픽스 I

수고하셨습니다!