RTong

(부제: R로 하는 통계분석방법)

지은이 : 알조장즈

(곽성화, 김수아, 정한수, 조건영)

학우분들의 통계분석 실력향상을 위해, 통계학과 4인방이 뭉쳤습니다.
자세한 코드 및 자료는 https://github.com/blueberrysmoooothie/RTong를 참고하세요!
※ '통'은 '뜻이 맞아 하나로 묶어진 무리'라는 뜻으로, R통은 **R로 묶어진 무리**라는 의미입니다.

통계분석

◎ 변수의 종류별 통계분석방법

독립변수		종속변수	분석방법	기타
~ ~	에 따른 에 대한	~ ~	의 차이는 없는가? 의 영향은 없는가?	
		수치형	일표본 평균검정	종속변수 1개
		수치형	대응표본 평균검정	종속변수 2개 (사전-사후)
		범주형	적합성검정(χ^2검정)	
범주형		수치형	독립표본 평균검정	독립변수 집단 2개 (남녀)
		수치형	일원 분산분석	독립변수 집단 3개 이상 (혈액형)
		수치형	이원 분산분석	범주형 독립변수 2개 (성별&혈액형)
		수치형	다변량 분산분석	수치형 종속변수 2개 이상
범주형		범주형	독립성검정(χ^2검정)	'교차분석'이라고도 함
수치형		수치형	상관분석	
		수치형	단순회귀분석	독립변수 1개 (직선)
		수치형	다항회귀분석	독립변수 1개 (곡선)
		수치형	다중회귀분석	독립변수 2개 이상
		수치형	더미변수 이용 회귀분석	
수치형		범주형	이항 로지스틱 회귀분석	종속변수 집단 2개 (생존-사망)
		범주형	다항 로지스틱 회귀분석	종속변수 집단 3개 이상

1. 일표본 평균검정(One sample t-test)

1) 양측검정

작년 동일 과목을 수강한 학생들의 나이는 평균 23살이었다.
올해 수강하는 학생들은 작년과 비교할 때 차이를 보이는지 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 학생들의 나이는 평균 23살이다. $(\mu = 23)$
- 대립가설(H_1) : 학생들의 나이는 평균 23살이 아니다. $(\mu \neq 23)$

```
library(MASS)
df <- na.omit(survey)
t.test(df$Age, mu=23)
```

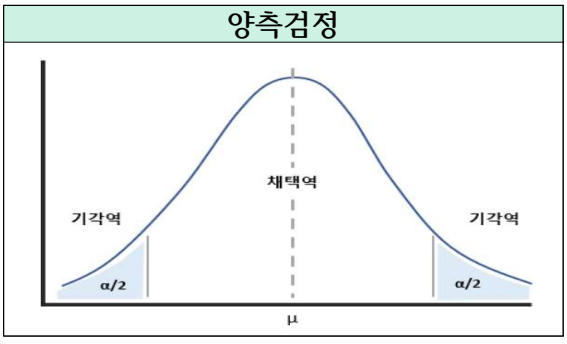
```
##  One Sample t-test
##
## data:  df$Age
## t = -5.4539, df = 167, p-value = 1.749e-07
## alternative hypothesis: true mean is not equal to 23
## 95 percent confidence interval:
##  19.50454 21.36261
## sample estimates:
## mean of x
##  20.43358
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(t) : -5.4539 / 유의확률($p-value$) : 1.749e-07
- 유의수준(α) : 0.05

③ 결과 해석

: '일표본 평균검정'의 결과 검정통계량 t=-5.4539이고,
p-value=0.0000001749이므로 유의수준 0.05에서 귀무가설이 기각되었다.
따라서 학생들의 나이는 평균 23살이 아니라고(차이를 보인다고) 할 수 있다.
올해 학생들의 나이는 평균 20.43358살로, 작년 수강한 학생들(평균 23살)과 비교할 때 어려졌다고 할 수 있다.



2) 단측검정(우측검정 alternative='greater')

작년 동일 과목을 수강한 학생들의 나이는 평균 21살이었다.
올해 수강하는 학생들은 작년과 비교할 때 차이를 보이는지 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 학생들의 나이는 평균 21살보다 작거나 같다. ($\mu \leq 21$)
- 대립가설(H_1) : 학생들의 나이는 평균 21살보다 크다. ($\mu > 21$)

```
t.test(df$Age, mu=21, alternative='greater')
```

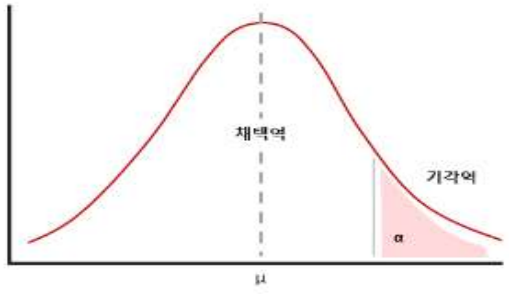
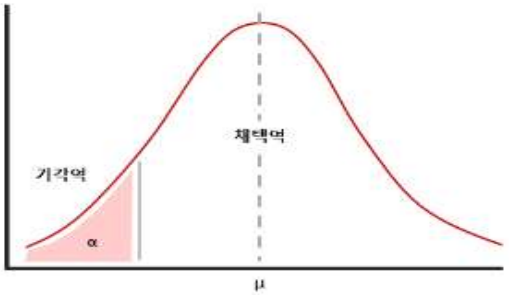
```
##  One Sample t-test
##
## data:  df$Age
## t = -1.2037, df = 167, p-value = 0.8848
## alternative hypothesis: true mean is greater than 21
## 95 percent confidence interval:
##  19.65524      Inf
## sample estimates:
## mean of x
##  20.43358
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(t) : -1.2037
- 유의확률($p-value$) : 0.8848
- 유의수준(α) : 0.05

③ 결과 해석

: '일표본 평균검정'의 결과 검정통계량 t=-1.2037이고,
p-value=0.8848이므로 유의수준 0.05에서 귀무가설이 채택되었다.
따라서 학생들의 나이는 평균 21살보다 작거나 같다고 할 수 있다.
올해 학생들의 나이는 평균 20.43358살이며, 95% 신뢰구간은
19.65524 ~ 무한대(Infinity)로 21살이 포함되어있다.

단측검정(우측)	단측검정(좌측)
	

3) 단측검정(좌측검정 alternative='less')

작년 동일 과목을 수강한 학생들의 나이는 평균 21살이었다.
올해 수강하는 학생들은 작년과 비교할 때 차이를 보이는지 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 학생들의 나이는 평균 21살보다 크거나 같다. ($\mu \geq 21$)
- 대립가설(H_1) : 학생들의 나이는 평균 21살보다 작다. ($\mu < 21$)

```
t.test(df$Age, mu=21, alternative='less')
```

```
##  One Sample t-test
##
## data:  df$Age
## t = -1.2037, df = 167, p-value = 0.1152
## alternative hypothesis: true mean is less than 21
## 95 percent confidence interval:
##      -Inf 21.21191
## sample estimates:
## mean of x
##  20.43358
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(t) : -1.2037
- 유의확률($p-value$) : 0.1152
- 유의수준(α) : 0.05

③ 결과 해석

: '일표본 평균검정'의 결과, 검정통계량 t=-1.2037이고
p-value=0.1152이므로 유의수준 0.05에서 귀무가설이 채택되었다.
따라서 학생들의 나이는 평균 21살보다 크거나 같다고 할 수 있다.
올해 학생들의 나이는 평균 20.43358살이며, 95% 신뢰구간은
-무한대(Infinity) ~ 21.21191로 21살이 포함되어있다.

2. 대응표본 평균검정(Paired t-test)

: 동일한 대상으로부터 사전과 사후 자료를 획득하여 이들의 차이를 비교,
또는 동일한 대상으로부터 두 번 조사가 된 경우 두 자료 간에 차이(왼손-오른손)
→ 대응되는 **자료 간의 차이를 일표본 자료로 취급**하므로 본질적으로 일표본 검정과 같음

동일한 대상으로부터 수면제 복용 여부(사전-사후)에 따른
수면시간(수치형 종속변수)의 차이를 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 수면제 복용여부에 따른 수면시간의 차이가 없다. ($\mu_1 - \mu_2 = 0$)
- 대립가설(H_1) : 수면제 복용여부에 따른 수면시간의 차이가 있다. ($\mu_1 - \mu_2 \neq 0$)

```
##  Paired t-test
##
## data:  extra by group
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean difference
##           -1.58
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(t) : -4.0621
- 유의확률($p-value$) : 0.002833
- 유의수준(α) : 0.05

③ 결과 해석

: '대응표본 평균검정'의 결과, 검정통계량 t=-4.0621이고,
p-value=0.002833이므로 유의수준 0.05에서 귀무가설이 기각되었다.
따라서 수면제 복용 여부에 따른 수면시간의 차이가 있다고 할 수 있다.
수면제 복용 전-후 수면시간의 평균 차이는 -1.58이다.
(복용 전 수면시간의 평균은 0.75이며, 복용 후 수면시간의 평균은 2.33이다.)

※ '대응표본 평균검정'의 경우 데이터 입력 방식에 따라 R코드 작성방법 다름
(롱 포맷 형태 vs 와이드 포맷 형태)

3. 독립표본 평균검정(Two sample t-test)

성별(범주가 2종류인 독립변수)에 따른 나이(수치형 종속변수)의 차이를 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 성별에 따른 평균 나이는 차이가 없다. ($\mu_{남} = \mu_{여}$)
- 대립가설(H_1) : 성별에 따른 평균 나이는 차이가 있다. ($\mu_{남} \neq \mu_{여}$)

```
##  Welch Two Sample t-test
##
## data:  Age by Sex
## t = -0.67983, df = 150.34, p-value = 0.4977
## alternative hypothesis: true difference in means between group Female and group Male
is not equal to 0
## 95 percent confidence interval:
##  -2.503417  1.221726
## sample estimates:
## mean in group Female   mean in group Male
##             20.11315             20.75400
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(t) : -0.67983
- 유의확률($p-value$) : 0.4977
- 유의수준(α) : 0.05

③ 결과 해석

: '독립표본 평균검정'의 결과, 검정통계량 t=-0.67983이고,
p-value=0.4977이므로 유의수준 0.05에서 귀무가설이 채택되었다.
따라서 학생들의 성별에 따른 나이 차이는 없다고 할 수 있다.
여성의 평균 나이는 20.11315이며, 남성의 평균 나이는 20.75400이다.

4. 일원분산분석(One-way ANOVA)

주로 사용하는 손 방향(범주가 3종류인 범주형 독립변수)에 따른 키(수치형 종속변수)의 차이를 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 손 방향에 따른 키의 차이가 없다. ($\mu_L = \mu_N = \mu_R$)
- 대립가설(H_1) : 적어도 하나의 손 방향에서는 키의 차이를 보인다.

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Clap          2     62   30.76   0.308  0.736
## Residuals   165  16504  100.02
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(F) : -0.308
- 유의확률($p-value$) : 0.736
- 유의수준(α) : 0.05

③ 결과 해석

: '일원분산분석'의 결과, 검정통계량 F=0.308이고, p-value=0.736이므로 유의수준 0.05에서 귀무가설이 채택되었다.
따라서 주로 사용하는 손 방향에 따른 키의 차이는 없다고 할 수 있다.
Tukey의 HSD방법에 의한 사후분석 결과, Neither과 Left 사이에서 가장 큰 차이를 보여줌을 알 수 있다.

◎ **사후분석(다중비교)**

: 각 집단들의 차이가 유의(H_0 기각)할 때, 차이가 어떤 집단들에게서 발생되고 있는지를 검토하기 위해 추가적인 통계분석을 실시하는 것

특징	사후검정 방법	특징
H0 채택을 꺼림 (진보적) ↑	Fisher의 LSD	ANOVA에서 F-검정이 유의한 경우에만 적용
	Duncan 방법	제1종 오류가 큰 단점, 반복수가 다른 경우에도 적용
	Tukey의 HSD	상당히 엄격한(보수적인) 검정
	Scheffe 방법	많이 사용되는 일반적인 방법

5. 이원분산분석(Two-way ANOVA)

주로 사용하는 손 방향(범주형 독립변수)과 성별(범주형 독립변수)에 따른 키(수치형 종속변수)의 차이를 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 모든 요인에 따른 키의 차이가 없다.
  (주효과와 상호작용효과가 없다)
- 대립가설(H_1) : 적어도 하나의 요인에서는 키의 차이를 보인다.
  (성별에 따른 키의 차이가 있다.
  적어도 하나의 손 방향에 따른 키의 차이가 있다.
  성별 및 손 방향의 상호작용 효과가 있다.)

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Sex          1   7938    7938 153.548 <2e-16 ***
## Clap         2    149      75   1.446  0.239
## Sex:Clap     2    104      52   1.002  0.370
## Residuals  162   8375      52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

② 검정통계량의 유의확률과 유의수준 비교

성별 - 검정통계량(F) : 153.548 / 유의확률($p-value$) : 0.0000000000000002

손 방향 - 검정통계량(F) : 1.446 / 유의확률($p-value$) : 0.239

상호작용효과 - 검정통계량(F) : 1.002 / 유의확률($p-value$) : 0.370

유의수준(α) : 0.05

③ 결과 해석

: '이원분산분석'의 결과, 유의수준 0.05에서 손 방향과 상호작용효과에 대한 귀무가설이 채택되었고, 유의수준 0.001에서 성별에 대한 귀무가설이 기각되었다.
따라서 성별에 따른 키의 차이는 있다고 할 수 있으며, 여성보다 남성의 키가 크다.

6. 교차분석(Chi-square test)

1) 독립성검정(independence test)

성별(범주형 독립변수)에 따른 흡연상태(범주형 종속변수)의 차이를 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 성별과 흡연상태는 서로 독립적이다(관련이 없다).
- 대립가설(H_1) : 성별과 흡연상태는 서로 독립적이지 않다(관련이 있다).

```
##  Pearson's Chi-squared test
##
## data:  Smoke and Sex
## X-squared = 5.0599, df = 3, p-value = 0.1675
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(χ^2) : 5.0599
- 유의확률($p-value$) : 0.1675
- 유의수준(α) : 0.05

③ 결과 해석

: '독립성검정'의 결과, 검정통계량 χ^2=5.0599이고, p-value=0.1675이므로 유의수준 0.05에서 귀무가설이 채택되었다.
따라서 성별과 흡연상태는 서로 독립적이라고 할 수 있다.

2) 적합성검정(Goodness of fit test)

범주형 변수가 하나일 경우, 모집단에서의 집단별 비율 분포를 검정하고자 한다.

① 가설 설정

- 귀무가설(H_0) : df 자료 내 남녀 비율이 3:2이다.
- 대립가설(H_1) : df 자료 내 남녀 비율이 3:2가 아니다.

```
##  Chi-squared test for given probabilities
##
## data:  table(df$Sex)
## X-squared = 7, df = 1, p-value = 0.008151
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(χ^2) : 7
- 유의확률($p-value$) : 0.008151
- 유의수준(α) : 0.05

③ 결과 해석

: '적합성검정'의 결과, 검정통계량 χ^2=7이고, p-value=0.008151이므로 유의수준 0.05에서 귀무가설이 기각되었다.
따라서 df 자료 내 남녀 비율이 3:2가 아니라고 할 수 있다.

7. 상관분석(Correlation analysis)

태양열, 풍속, 온도(수치형 독립변수)가 Ozone(수치형 종속변수)에 영향을 주는지(상관관계가 있는지)를 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 태양열 및 풍속, 온도와 오존량은 관계가 없다.
- 대립가설(H_1) : 태양열 및 풍속, 온도과 오존량은 관계가 있다.

```
## Call:corr.test(x = df)
## Correlation matrix
##         Ozone Solar.R  Wind  Temp Month   Day
## Ozone    1.00    0.35 -0.61  0.70  0.14 -0.01
## Solar.R  0.35    1.00 -0.13  0.29 -0.07 -0.06
## Wind    -0.61   -0.13  1.00 -0.50 -0.19  0.05
## Temp     0.70    0.29 -0.50  1.00  0.40 -0.10
## Month    0.14   -0.07 -0.19  0.40  1.00 -0.01
## Day     -0.01   -0.06  0.05 -0.10 -0.01  1.00
## Sample Size
## [1] 111
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##         Ozone Solar.R Wind Temp Month Day
## Ozone    0.00    0.00 0.00 0.00  1.00   1
## Solar.R  0.00    0.00 1.00 0.02  1.00   1
## Wind     0.00    0.18 0.00 0.00  0.37   1
## Temp     0.00    0.00 0.00 0.00  0.00   1
## Month    0.13    0.44 0.04 0.00  0.00   1
## Day      0.96    0.55 0.60 0.31  0.93   0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

② 검정통계량의 유의확률과 유의수준 비교

태양열과 오존 - 검정통계량(r) : 0.35 / 유의확률($p-value$) : 0.00

풍속과 오존 - 검정통계량(r) : -0.61 / 유의확률($p-value$) : 0.00

온도과 오존 - 검정통계량(r) : 0.70 / 유의확률($p-value$) : 0.00

유의수준(α) : 0.05

③ 결과 해석

: '상관분석'의 결과, 태양열 및 풍속, 온도와 오존량은 각 p-value가 0.00이므로 유의수준 0.05에서 귀무가설이 기각되었다.

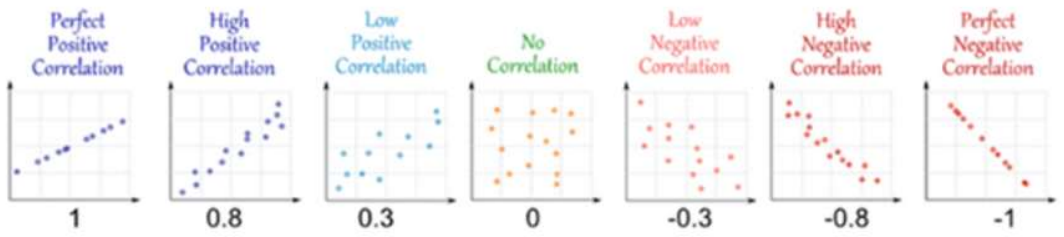
따라서 태양열 및 풍속, 온도와 오존량은 관계가 있다고 할 수 있다.

온도와 오존량 간의 상관계수r=0.70으로 강한 양의 상관관계,

풍속과 오존량 간의 상관계수r=-0.61로 강한 음의 상관관계,

태양열과 오존량 간의 상관계수r=0.35로 보통 양의 상관관계가 있다.

◎ 상관계수 해석



0.0 r 0.1	거의 관계 없음
0.1 ~ 0.2	약한 양의 상관관계
0.2 ~ 0.4	보통의 양의 상관관계
0.4 ~ 0.6	비교적 강한 양의 상관관계
0.6 ~ 0.8	강한 양의 상관관계
0.8 ~ 1.0	매우 강한 양의 상관관계

8. 단순회귀분석(Simple linear regression analysis)

기온(수치형 독립변수)이 오존량(수치형 종속변수)에 미치는 영향을 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 회귀식은 유의하지 않다(회귀식의 회귀계수는 0이다).
- 대립가설(H_1) : 회귀식은 유의하다(회귀식의 회귀계수는 0이 아니다).

```
## Call:
## lm(formula = Ozone ~ Temp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.922 -17.459  -0.874  10.444 118.078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -147.6461    18.7553  -7.872 2.76e-12 ***
## Temp           2.4391     0.2393  10.192  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.92 on 109 degrees of freedom
## Multiple R-squared:  0.488,  Adjusted R-squared:  0.4833
## F-statistic: 103.9 on 1 and 109 DF,  p-value: < 2.2e-16
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(F) : 103.9
- 유의확률($p-value$) : 2.2e-16
- 유의수준(α) : 0.05

③ 결과 해석

: '단순회귀분석'의 결과, 검정통계량 F=103.9이고, p-value=0.0000000000000022이므로 유의수준 0.05에서 귀무가설이 기각되었다. 따라서 회귀식은 통계적으로 유의하다고 할 수 있다.

결정계수 R^2=0.488이므로, 48.8%의 회귀모형 설명력을 가진다고 할 수 있다.

- 회귀식 : 추정된 오존량($\hat{Y}$) = -147.6461 + 2.4391 × Temp(X_1)

온도가 올라갈수록 오존량은 높아지는 경향을 보이고 있었다.

10. 다중회귀분석(Multiple linear regression analysis)

태양열 및 풍속, 온도, 월(수치형 독립변수 4종)이 오존량(수치형 종속변수)에 미치는 영향을 알아보고자 한다.

① 가설 설정

- 귀무가설(H_0) : 회귀식은 유의하지 않다(회귀식의 회귀계수는 0이다).
- 대립가설(H_1) : 회귀식은 유의하다(회귀식의 회귀계수는 0이 아니다).

```
## Call:
## lm(formula = Ozone ~ Temp + Wind + Solar.R + Month, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.870 -13.968  -2.671   9.553  97.918
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -58.05384   22.97114  -2.527   0.0130 *
## Temp          1.87087    0.27363   6.837 5.34e-10 ***
## Wind         -3.31651    0.64579  -5.136 1.29e-06 ***
## Solar.R       0.04960    0.02346   2.114   0.0368 *
## Month        -2.99163    1.51592  -1.973   0.0510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.9 on 106 degrees of freedom
## Multiple R-squared:  0.6199, Adjusted R-squared:  0.6055
## F-statistic: 43.21 on 4 and 106 DF,  p-value: < 2.2e-16
```

② 검정통계량의 유의확률과 유의수준 비교

- 검정통계량(F) : 43.21
- 유의확률($p-value$) : 2.2e-16
- 유의수준(α) : 0.05

③ 결과 해석

: '다중회귀분석'의 결과, 검정통계량 F=43.21이고, p-value=0.0000000000000022이므로 유의수준 0.05에서 귀무가설이 기각되었다. 따라서 회귀식은 통계적으로 유의하다고 할 수 있다.

수정결정계수 R^2=0.6055이므로, 회귀모형 설명력은 약 60%로 높은

수준이라고 할 수 있다. (독립변수의 수가 많아짐에 따라 결정계수도 커지는 단점을 보완하고자 수정된 결정계수를 확인함)

* 회귀식 : 추정된 오존량($\hat{Y}$) = -58.05384 + 1.87087 × Temp(X_1) + -3.31651 × Wind(X_2) + 0.04960 × Solar.R(X_3) - 2.99163 × Month(X_4)

온도, 풍속, 태양열은 오존량에 유의한 영향이 되고 있었다. 온도와 태양열은 증가할수록 오존량이 증가하였으나, 풍속은 증가할수록 오존량이 낮아지는 경향을 보인다.

세 요인의 영향력의 크기는 표준화계수(β)로 비교가 가능한데, 절대값을 기준으로 '풍속'의 영향력(-3.31651)이 '온도'(1.87087)와 '태양열'(0.04960)의 영향력에 비해 크게 나타났다.

◎ 유의한 독립변수의 선택방법

전진선택법	독립변수를 하나씩 추가해나가면서 관계의 유의성 여부를 판단하는 방법. 유의하지 않은 변수가 나타날 때까지 독립변수의 추가를 지속함
후방소거법	전진선택법과 반대 방식 모든 독립변수가 투입된 상태에서 회귀분석을 실시한 후, 유의확률이 가장 큰 독립변수의 순으로 차례로 제거해나가면서 유의하지 않은 변수가 더 이상 없을 때까지 반복적으로 실시함
단계선택법	전진선택법을 기본으로 하여 후방소거법을 조합한 방법 가장 많이 쓰이는 변수선택 방법

11. 이항 로지스틱 회귀분석(Binomial logistic regression analysis)

결과변수(종속변수)가 이분형 범주(예/아니오, 성공/실패, 생존/사망 등)를 가질 때 예측변수(독립변수)로부터 결과변수의 범주를 예측한다.
결과변수는 항상 0과 1 사이의 확률값을 가지며, 일정 기준값(0.5) 보다 크면 사건이 발생한 것으로 예측함

① 가설 설정

- 귀무가설(H_0) : 회귀식은 유의하지 않다(회귀모델이 적합하지 않다)
  or 회귀식의 회귀계수는 0이다(회귀계수가 의미가 없다)
- 대립가설(H_1) : 회귀식은 유의하다(회귀모델이 적합하다)
  or 회귀식의 회귀계수는 0이 아니다(회귀계수가 의미가 있다)

교수님 수업자료 Ch.16 p15~p27 예제 참고

```
churn.logit <- glm(churn~ . , data=churn.train, family=binomial(link="logit"))
summary(churn.logit)
```

```
Call:
glm(formula = churn ~ ., family = binomial(link = "logit"), data = churn.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1532  -0.5132  -0.3402  -0.1953   3.2528

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -8.6515638  0.7243142 -11.944  < 2e-16 ***
account_length              0.0008458  0.0013912   0.608 0.543199
international_planyes       2.0427543  0.1454974  14.040  < 2e-16 ***
voice_mail_planyes         -2.0250146  0.5740840  -3.527 0.000420 ***
number_vmail_messages       0.0358803  0.0180108   1.992 0.046355 *
total_eve_calls             0.0010579  0.0027826   0.380 0.703817
total_eve_charge           -9.5463678 19.2437266  -0.496 0.619840
total_night_minutes        -0.1238287  0.8764906  -0.141 0.887650
total_night_calls           0.0006993  0.0028419   0.246 0.805628
total_night_charge          2.8338084 19.4769043   0.145 0.884319
total_intl_minutes         -4.3377914  5.3009719  -0.818 0.413185
total_intl_calls           -0.0929680  0.0250603  -3.710 0.000207 ***
total_intl_charge          16.3900316 19.6323938   0.835 0.403804
number_customer_service_calls  0.5135638  0.0392678  13.079  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2758.3  on 3332  degrees of freedom
Residual deviance: 2158.7  on 3315  degrees of freedom
AIC: 2194.7

Number of Fisher Scoring iterations: 6
```

②-1 결과 해석

: '이항 로지스틱 회귀분석'의 결과, international_planyes와 voice_mail_planyes, number_vmail_messages, total_intl_calls, number_customer_service_calls 변수들의 p값은 유의수준 0.05보다 작으므로 귀무가설이 기각되어, 회귀계수가 의미가 있다고 할 수 있다.

voice_mail_planyes, total_intl_calls은 증가할수록 고객이탈은 낮아지고, international_planyes와 number_vmail_messages, number_customer_service_calls은 증가할수록 고객이탈은 높은 것으로 나타났다.

영향력의 크기는 표준화계수(β)로 비교가 가능한데, 절대값을 기준으로 international_planyes(2.0427543), voice_mail_planyes(-2.0250146), number_customer_service_calls (0.5135638), total_intl_calls(-0.0929680), number_vmail_messages(0.0358803) 순으로 영향력에 크게 나타났다.

```
# 오즈(Odds 계산)
exp(coef(churn.logit))
```

```
                 (Intercept)                account_length
                1.748532e-04                  1.000846e+00
       international_planyes                voice_mail_planyes
                7.711821e+00                  1.319919e-01
      number_vmail_messages                total_day_minutes
                1.036532e+00                  7.833315e-01
           total_day_calls                   total_day_charge
                1.003201e+00                  4.539006e+00
         total_eve_minutes                   total_eve_calls
                2.267538e+00                  1.001058e+00
          total_eve_charge                 total_night_minutes
                7.146035e-05                  8.835312e-01
        total_night_calls                   total_night_charge
                1.000700e+00                  1.701012e+01
...(이하 생략)
```

②-2 결과 해석

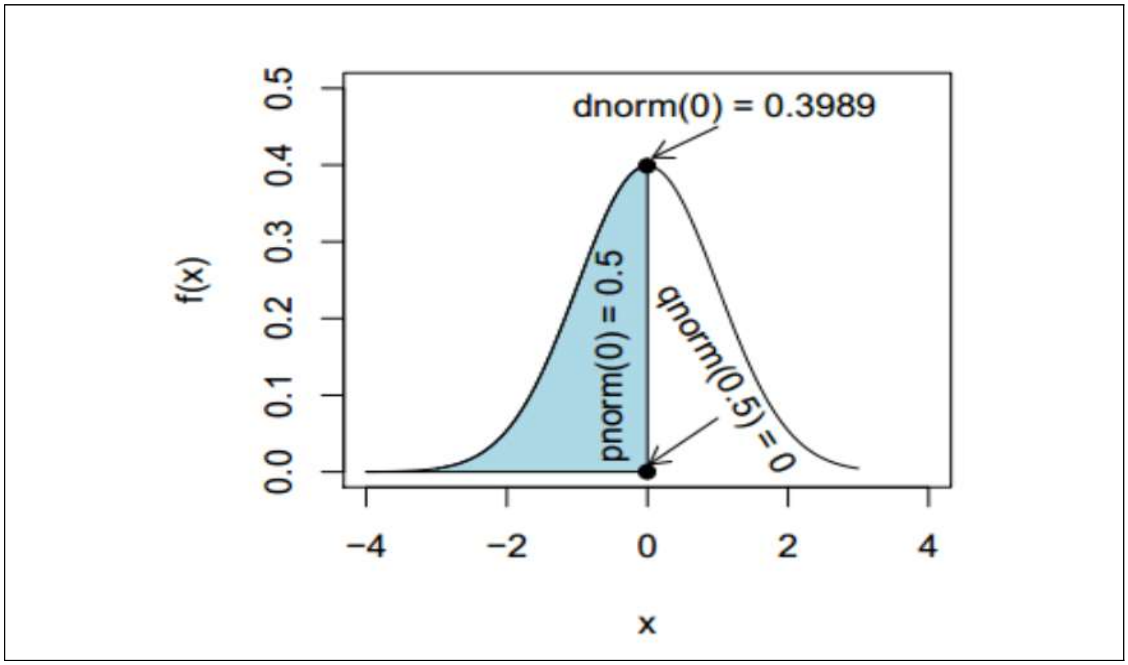
: [오즈비는 해당 독립변수의 값이 1 증가하면 종속변수가 발생할 확률 증가를 의미하며, 회귀계수가 의미 있는 것($p-value$ < 유의수준α)만 읽어주면 됨]

international_planyes 값이 1 증가하면 이탈확률이 약 7.7배 증가하며, voice_mail_planyes 값이 1 증가하면 약 1.3배 증가한다.

③ 최종 결론

: 이동통신회사의 고객이탈에 영향을 주는 변수로는 international_planyes와 voice_mail_planyes, number_vmail_messages, total_intl_calls, number_customer_service_calls로 나타났다.
voice_mail_planyes, total_intl_calls은 증가할수록 고객이탈은 낮아지고, international_planyes와 number_vmail_messages, number_customer_service_calls은 증가할수록 고객이탈은 높은 것으로 나타났다.
이중에서 가장 영향을 주는 요인으로는 international_planyes(β=2.0427543)로 나타났으며, international_planyes 값이 1 증가하면 고객이탈확률이 7.7배 증가하는 것으로 나타났다.

◎ 확률분포 관련 함수



확률밀도함수(PDF)	백분위수함수	확률분포함수
x에 x값을 넣으면 f(x)값을 알려줌 해당 값이 나올 확률	q에 x의 값을 넣으면 F(x)값(면적)을 알려줌 해당 값까지의 누적확률값	p에 F(x)의 값(면적)을 넣으면 x값을 알려줌
dnorm(x=0) ▶ 0.3989423 dnorm(x=2) ▶ 0.05399097	pnorm(q=1.65) ▶ 0.9505285 (약 0.95) pnorm(q=-1.65) ▶ 0.04947147 (약 0.05)	qnorm(p=0.95) ▶ 1.644854 qnorm(p=0.05) ▶ -1.644854

구분	이항분포	정규분포	t분포	F분포	χ^2분포	일양분포
확률밀도함수	dbinom()	dnorm()	dt()	df()	dchisq()	dunif()
확률분포함수	qbinom()	qnorm()	qt()	qf()	qchisq()	qunif()
백분위수함수	pbinom()	pnorm()	pt()	pf()	pchisq()	punif()
난수생성함수	rbinom()	rnorm()	rt()	rf()	rchisq()	runif()

◎ 데이터 요약

1) 범주형 변수

① 중심경향치(central tendency)

구분	설명	R 코드
최빈값(mode) 최빈수, 최빈치	빈도가 가장 많은 관측값 자료에서 반드시 하나만 존재X	최빈값에 대한 함수는 없음

2) 수치형 변수

① 중심경향치(central tendency)

구분	설명	R 코드
평균(mean) 산술평균	관측값들의 합을 관측값의 개수로 나눈 값 극단값(이상치) 영향을 많이 받음	mean()
중앙값(median) 중앙치, 중위수	자료를 크기순으로 나열했을 때 중앙에 위치하는 값	median()

② 산포도(degree of scttering)

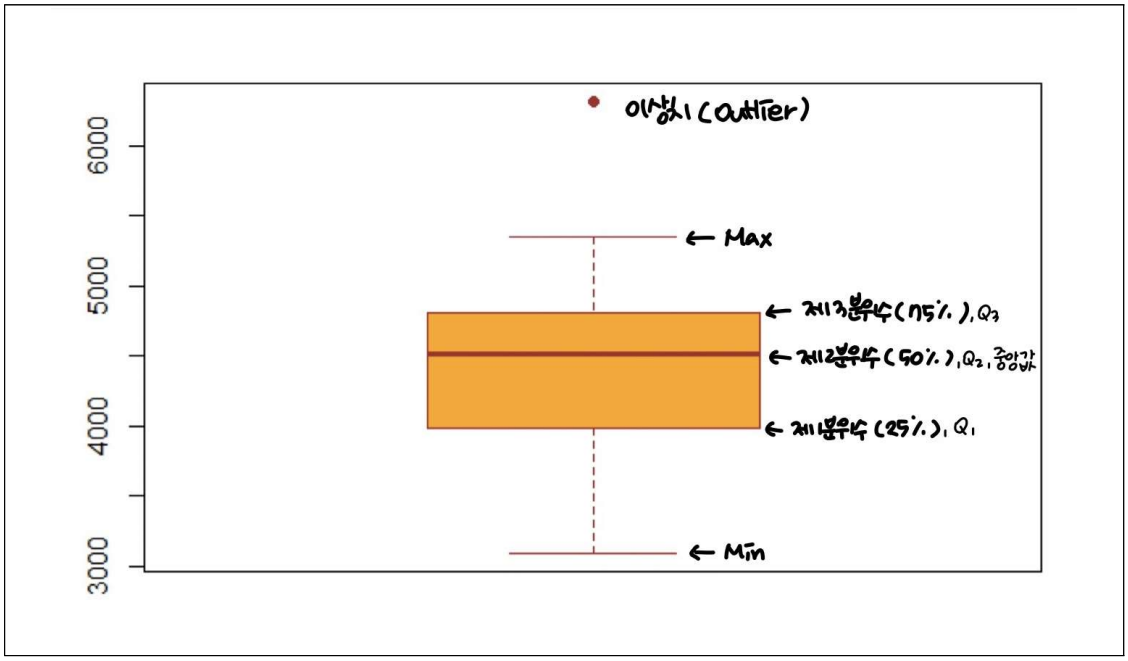
구분	설명	R 코드
범위 (range)	최대값 - 최소값	최댓값 : max() 최솟값 : min()
사분위수 범위 (InterQuartile Range)	자료를 오름차순으로 정렬했을 때, 상위 25%와 하위 25%를 제외하고 범위를 구한 값. 극단값의 영향을 크게 받지 않음 $IQR = Q_3 - Q_1$	사분위수 : quantile()
사분위수 편차 (interquartile range)	오름차순으로 정렬했을 때, 3사분위수와 1사분위수의 평균 $Q = \frac{Q_3 - Q_1}{2}$	사분위수 : quantile()
분산 (variance)	관측값들이 평균으로부터 얼마나 떨어져 있는 정도(거리) 음의 거리를 배제하기 위해 제곱 사용 $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{n-1}$	분산 : var()
표준편차 (standard deviation)	분산의 양의 제곱근 관측치의 단위와 동일하므로 두 집단을 상대적으로 비교 가능	표준편차 : sd()

※ 편차(deviation) : 실제 데이터의 값(관측값)에서 표본평균을 차감한 것. 편차의 합은 0

※ 오차(error) : 측정대상의 참값과 측정도구를 통한 측정값 사이의 불일치 정도
모집단으로부터 추정한 회귀식으로부터 얻은 예측값과 실제 관측값의 차이

※ 잔차(residual) : 표본으로 추정한 회귀식의 값(예측값)과 실제 데이터의 값(관측값)의 차이

◎ 상자그림



[출처]
https://dlearner.tistory.com/36
https://dogmas.tistory.com/entry/R에서-정규분포-dnorm-pnorm-qnorm-rnorm-함수이용
https://muzukphysics.tistory.com/87