

쉽게 배우는 통계입문 (2)

I n t r o d u c t i o n t o S t a t i s t i c s

Present by Hobin Kwak

쉽게 배우는 통계입문의 심화 버전!
(중복되는 내용 있음! 그러나, 듣길 추천)

[강의계획]

- 표본분포 (쉽게 배우는 통계입문 복습)
- 통계적 추론
- 검정
- 회귀분석
- 분산분석
- 범주형 자료 분석
- 비모수적 추론

표본분포

1. 표본 분포 (sampling distribution)

- : 모집단에서 일정 크기로 표본을 뽑을 때, 그 표본의 통계량의 확률분포
- : 통계적 추정/검정의 핵심
- : 예시) 아래와 같은 분포를 띤 모집단에서 크기가 2인 확률표본 X_1 과 X_2 를 추출할 때, 표본평균의 확률분포는?

| | | |
|----------|-----|-----|
| X | 0 | 1 |
| $f(X=x)$ | 0.3 | 0.7 |

| | | |
|---|------|------|
| | 0 | 1 |
| 0 | 0.09 | 0.21 |
| 1 | 0.21 | 0.49 |

표본분포

1. 표본평균의 평균과 표준편차

: X_1, \dots, X_n 이 모평균 μ , 모표준편차 σ 인
모집단으로부터의 확률표본 (i.i.d)일 때,

표본평균 $\bar{X} = \frac{\sum X_i}{n}$

$$E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}[E(X_1) + \dots + E(X_n)] = \frac{1}{n}n\mu = \mu$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

표본분포 - 중심극한정리

1. 중심극한정리

: 평균이 μ , 표준편차 σ 인 임의의 모집단으로부터
크기 n 인 표본에서의 **표본평균**은 n (주로 30 이상)이 크면 근사적으로
평균이 μ 이고 분산이 $\frac{\sigma^2}{n}$ 인 정규분포를 따름

: 모집단이 어떤 형태의 분포든 **표본 크기가 크면 항상 성립**
: **모집단이 정규분포라면** 표본평균은 표본 개수와
상관없이 **항상 정규분포**를 따른다.

2. 예) 어느 회사 그래픽카드의 평균수명은 3년이고 표준편차는 2년이다.
이 회사 제품에서 100개의 그래픽카드를 뽑아 평균 수명을 확인할 때
4년 이상일 확률은?

표본분포 - 중심극한정리

1. 이항분포의 정규분포 근사

: 서로 독립이고 동일한 모수 p 를 갖는 베르누이 확률변수

Y_1, Y_2, \dots, Y_n 에 대해 $X = Y_1 + Y_2 + \dots + Y_n$

Y_1, Y_2, \dots, Y_n 로부터의 표본평균에 대해 중심극한정리 적용

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\sum_{i=1}^n Y_i - np}{\sqrt{np(1-p)}} = \frac{\bar{Y} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1^2)$$

2. 표본비율 정규근사

: 베르누이분포로부터의 크기 n 인 확률표본에 대해,
표본비율 \hat{p} 의 분포는 n 이 클 때, 근사적으로

$$N\left(p, \frac{p(1-p)}{n}\right)$$

표본분포 - 카이제곱분포

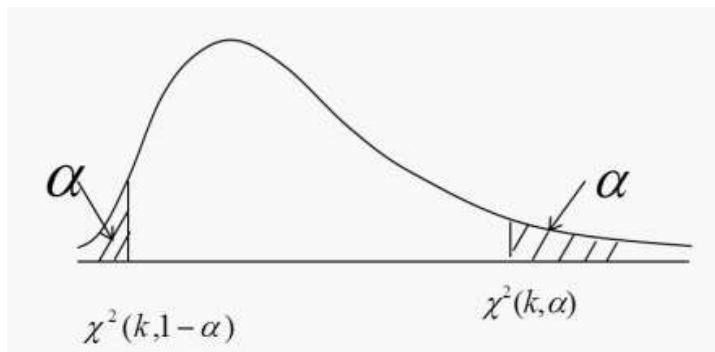
1. 카이제곱(χ^2) 분포

: 표본분산과 관련된 분포

: 확률변수 Z_1, \dots, Z_k 가 각각 표준정규분포를 따르고 독립일 때 그들의 제곱합은 자유도 k 인 카이제곱 분포 $\chi^2_{(k)}$ 를 따름

$$Z_1^2 + Z_2^2 + \dots + Z_k^2 \sim \chi^2_{(k)}$$

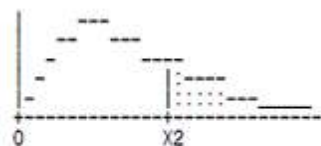
: 표본분산(s^2)을 알고 모분산(σ^2)을 추정할 때 사용하는 분포
(표본크기 클 수록 치우침이 적어짐)



표본분포 - 카이제곱분포

1. 카이제곱(χ^2) 분포표

CHI-SQUARE TABLE: VALUES OF CHI-SQUARE (ALPHA) OF THE CHI-SQUARE DISTRIBUTION



| DF | X ² (.995) | X ² (.99) | X ² (.975) | X ² (.95) | X ² (.90) | X ² (.05) | X ² (.025) | X ² (.01) | X ² (.005) |
|----|-----------------------|----------------------|-----------------------|----------------------|----------------------|----------------------|-----------------------|----------------------|-----------------------|
| 1 | 0.000 | 0.000 | 0.001 | 0.004 | 3.841 | 5.024 | 6.635 | 7.879 | |
| 2 | 0.010 | 0.020 | 0.051 | 0.103 | 5.991 | 7.378 | 9.210 | 10.597 | |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 7.815 | 9.348 | 11.345 | 12.838 | |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 9.488 | 11.143 | 13.277 | 14.860 | |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 11.071 | 12.833 | 15.086 | 16.750 | |
| 6 | 0.676 | 0.872 | 1.237 | 1.635 | 12.592 | 14.449 | 16.812 | 18.548 | |
| 7 | 0.989 | 1.239 | 1.690 | 2.167 | 14.067 | 16.013 | 18.475 | 20.278 | |
| 8 | 1.344 | 1.646 | 2.180 | 2.733 | 15.507 | 17.535 | 20.090 | 21.955 | |
| 9 | 1.735 | 2.088 | 2.700 | 3.325 | 16.919 | 19.023 | 21.666 | 23.589 | |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 18.307 | 20.483 | 23.209 | 25.188 | |
| 11 | 2.603 | 3.053 | 3.816 | 4.575 | 19.675 | 21.920 | 24.725 | 26.757 | |
| 12 | 3.074 | 3.571 | 4.404 | 5.226 | 21.026 | 23.337 | 26.217 | 28.300 | |
| 13 | 3.565 | 4.107 | 5.009 | 5.892 | 22.362 | 24.736 | 27.688 | 29.819 | |
| 14 | 4.075 | 4.660 | 5.629 | 6.571 | 23.685 | 26.119 | 29.141 | 31.319 | |
| 15 | 4.601 | 5.229 | 6.262 | 7.261 | 24.996 | 27.488 | 30.578 | 32.801 | |
| 16 | 5.142 | 5.812 | 6.908 | 7.962 | 26.296 | 28.845 | 32.000 | 34.267 | |
| 17 | 5.697 | 6.408 | 7.564 | 8.672 | 27.587 | 30.191 | 33.409 | 35.718 | |
| 18 | 6.265 | 7.015 | 8.231 | 9.390 | 28.869 | 31.526 | 34.805 | 37.156 | |
| 19 | 6.844 | 7.633 | 8.907 | 10.117 | 30.144 | 32.852 | 36.191 | 38.582 | |
| 20 | 7.434 | 8.260 | 9.591 | 10.851 | 31.410 | 34.170 | 37.566 | 39.997 | |
| 21 | 8.034 | 8.897 | 10.283 | 11.591 | 32.671 | 35.479 | 38.932 | 41.401 | |
| 22 | 8.643 | 9.542 | 10.982 | 12.338 | 33.924 | 36.781 | 40.289 | 42.796 | |
| 23 | 9.260 | 10.196 | 11.689 | 13.091 | 35.172 | 38.076 | 41.638 | 44.181 | |
| 24 | 9.886 | 10.856 | 12.401 | 13.848 | 36.415 | 39.364 | 42.980 | 45.559 | |
| 25 | 10.520 | 11.524 | 13.120 | 14.611 | 37.652 | 40.646 | 44.314 | 46.928 | |
| 26 | 11.160 | 12.198 | 13.844 | 15.379 | 38.885 | 41.923 | 45.642 | 48.290 | |
| 27 | 11.808 | 12.879 | 14.573 | 16.151 | 40.113 | 43.195 | 46.953 | 49.645 | |
| 28 | 12.461 | 13.565 | 15.308 | 16.928 | 41.337 | 44.461 | 48.278 | 50.993 | |
| 29 | 13.121 | 14.256 | 16.047 | 17.708 | 42.557 | 45.722 | 49.588 | 52.336 | |
| 30 | 13.787 | 14.953 | 16.791 | 18.493 | 43.773 | 46.979 | 50.892 | 53.672 | |

표본분포 - 카이제곱분포

1. 정규모집단에서의 표본분산 분포

: X_1, \dots, X_n 을 정규분포 $N(\mu, \sigma^2)$ 으로부터의 확률표본이라 할 때,
표본분산 $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

2. 증명

$\frac{X_i - \mu}{\sigma} \sim N(0,1)$ 이고 서로 독립($i = 1, 2, \dots, n$)

$$\sum \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum \left(\frac{(X_i - \bar{X})}{\sigma} \right)^2 + n \left(\frac{(\bar{X} - \mu)}{\sigma} \right)^2$$

$$= \left(\frac{(n-1)S^2}{\sigma^2} \right) + \left(\frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}} \right)^2$$

표본분포 - 카이제곱분포

1. 분산이 동일한 두 정규모집단에서의 표본분산의 분포

: X_1, \dots, X_n 과 Y_1, \dots, Y_n 가 각각 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)$ 을 따르며 서로 독립인 확률표본이라 할 때, 표본분산은 각각

$$S_1^2 = \frac{\sum_i (X_i - \bar{X})^2}{n_1 - 1}, \quad S_2^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n_2 - 1} \quad \text{일 경우}$$

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

2. 합동표본분산(pooled sample variance)

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$

표본분포 - 카이제곱분포

1. 단봉분포
2. 오른쪽에 꼬리를 가짐
3. 항상 양수값을 가짐
4. $E(\chi_\phi^2) = \phi, \text{Var}(\chi_\phi^2) = 2\phi$
5. 자유도가 커지면 정규분포에 가까워짐
6. 모분산 추정 및 검정에 활용
7. 적합성, 동질성, 독립성 검정 등에 사용

표본분포 - t분포

1. t분포

- : X의 분포가 정규분포일 때, 표본평균의 분포에서 모집단의 표준편차를 모를 경우
모표준편차(σ) 대신 표본표준편차(s)를 사용
- : t분포는 자유도에 의해 모양이 결정됨
 - 자유도: 임의로 결정될 수 있는 수
- : $Z \sim N(0,1)$, $V \sim \chi^2_{(k)}$ 이고 Z와 V는 서로 독립일 때,

$$T = \frac{Z}{\sqrt{V/k}} \sim t(k)$$

- : $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ 일 때,

$$t(n-1) \sim \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

표본분포 - t분포

1. 분산이 동일한 두 정규모집단에서의 t-분포

: X_1, \dots, X_n 과 Y_1, \dots, Y_n 가 각각 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ 을 따르며 서로 독립인 확률표본이라 할 때, 표본분산은 각각

$$S_1^2 = \frac{\sum_i (X_i - \bar{X})^2}{n_1 - 1}, \quad S_2^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n_2 - 1} \text{ 일 경우}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p(\sqrt{1/n_1 + 1/n_2})} \sim t(n_1 + n_2 - 2)$$

표본분포 - t분포

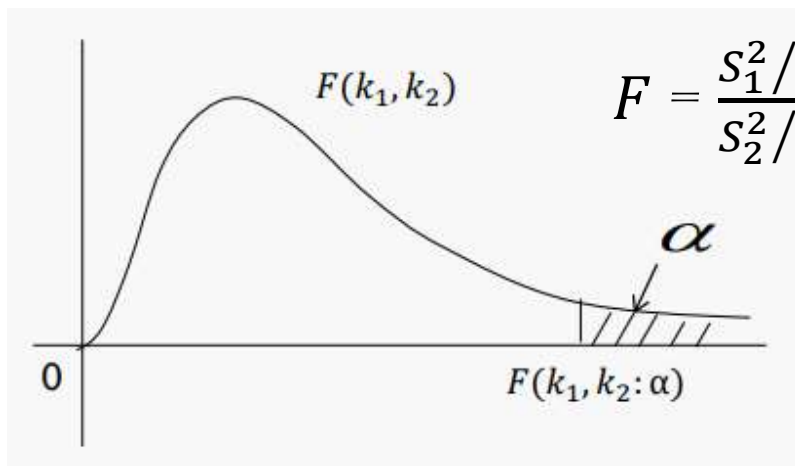
1. T분포는 정규분포보다 넓게 퍼져 있고 꼬리부분이 더 평평함
2. Bell Shaped
3. 표본크기가 커질수록 분포가 중심부근에서 점점 더 뾰족해짐
 - 표본 크기가 30 이상이 되면 **정규분포에 근사**
4. 주로 모평균 추정 혹은 모평균차이에 대한 추정 시
모표준편차를 모를 때 t분포를 사용함
5. 표본 크기가 30 이상일 경우에는 표준정규분포, 미만일 때는 t분포

표본분포 - F-분포

1. F분포

- : F-분포는 두 정규모집단의 분산을 비교하는 추론에 사용
- : V_1 과 V_2 는 각각 자유도 k_1, k_2 인 카이제곱분포를 따르는 독립인 확률변수

$$F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2) \quad \frac{1}{F} = \frac{V_2/k_2}{V_1/k_1} \sim F(k_2, k_1)$$



$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

표본분포 - F-분포

1. F-분포와 t-분포와의 관계

: $Z \sim N(0,1), V \sim \chi^2(k)$ 이고 Z 와 V 는 서로 독립일 때,

$$T = \frac{Z}{\sqrt{\frac{V}{k}}} \sim t(k), \quad T^2 = \frac{Z^2/1}{\frac{V}{k}} \sim F(1, k)$$

표본분포 - 정리

1. 정규분포

- 모분산을 알고 있을 때, 모평균에 대한 추정/검정
- 모분산을 알고 있을 때, 두 모평균 차이에 대한 추정/검정
- 표본크기가 클 때 모평균 혹은 모평균 차이에 대한 추정/검정
- 표본크기가 클 때, 모비율 혹은 모비율 차이에 대한 추정/검정

2. t 분포

- 모분산을 모를 때 모평균에 대한 추정/검정
- 모분산을 모를 때 두 모평균 차이에 대한 추정/검정

3. 카이제곱분포

- 모분산에 대한 추정/검정
- 분할표에 의한 독립성/적합성/동질성 검정

4. F 분포

- 두 모분산 차이에 대한 추정/검정
- 분산분석표의 요인에 관한 추정/검정

1. 통계적 추정(estimation)

- : 표본의 통계량을 기초로 하여
모집단의 모수를 추정하는 방법론
- : 통계적 추정의 관심대상은 **통계량이 아니라 모수**

2. 통계적 추정의 종류

1. 점 추정 (Point estimation)

- 모수를 단일한 값으로 추측하는 방식
- 신뢰도를 나타낼 수 없음 (추정치가 어느정도 옳을지 모름)
- 오차에 대한 정보가 없음

2. 구간추정

- 모수를 포함한다고 추정되는 구간을 구하는 방식
- 신뢰도를 나타낼 수 있음

추정 - 점 추정

1. 추정량 (estimator)

- : 확률변수 (표본 추출 전에는 알 수 없음)
- : 모수에 대한 대략적 정보 제공

2. 추정값 (estimate)

- : 표본 추출로 추정량을 통해 실현된 값

3. 통계량과 추정량의 차이

- : 추정량 = 추정에 사용되는 통계량
- : 추정값 = 표본에서 결정된 추정량의 값

추정 - 점 추정

1. 불편성 (Unbiasedness)

: 모수의 추정량의 기댓값이 모수가 되는 성질

2. 유효성 (Efficiency)

: 추정량이 불편추정량이고 분산이
다른 추정량에 비해 가장 작은 분산을 갖는 성질

3. 일치성 (Consistency)

: 표본 크기가 커질 수록 추정량이 모수에 수렴하는 성질

4. 충분성 (Sufficiency)

: 모수에 대해 가능한 많은 표본정보를 내포하는 성질

추정 - 점 추정

1. 불편성 (Unbiasedness)

: 모든 모수 θ 의 모든 참값에 대하여 $E(\hat{\theta}) = \theta$ 이면 $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ 을 모수의 불편추정량이라고 함

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

2, 유효성

: 추정량의 표준오차(standard error): 흩어짐의 정도를 나타내는 척도, 추정량 $\hat{\theta}$ 의 표준편차

두 불편추정량 $\hat{\theta}_1, \hat{\theta}_2$ 에 대해 $S.E.(\hat{\theta}_1) < S.E.(\hat{\theta}_2)$ 이면
추정량 $\hat{\theta}_1$ 이 추정량 $\hat{\theta}_2$ 보다 유효

추정 - 점 추정

1. 모평균의 추정

: 추정량: 표본평균 $\hat{\mu} = \bar{X}$

: 표준오차: S.E. ($\hat{\mu}$) = $\frac{\sigma}{\sqrt{n}}$

: 표준오차의 추정량: S.E. ($\widehat{\hat{\mu}}$) = $\frac{s}{\sqrt{n}}$

2, 오차한계(limit of error)

: $P\left\{|\bar{X} - \mu| \leq \frac{2\sigma}{\sqrt{n}}\right\} = 0.954$

: 추정량 $\hat{\mu} = \bar{X}$ 을 사용하여 모평균을 1,000번 추정하면
오차가 $\frac{2\sigma}{\sqrt{n}}$ 이내인 것인 대략 954번 (정규모집단의 경우)

-> $\frac{2\sigma}{\sqrt{n}}$ 는 $\hat{\mu} = \bar{X}$ 의 95.4% (근사)오차한계

추정 - 점 추정

1. 모비율의 추정

: $X \sim B(n, p)$: $E(X) = np$, $Var(X) = npq$

: 표본비율 $\hat{p} = \frac{X}{n}$ 은 모비율 p 의 불편추정량

$$E(\hat{p}) = p, \quad S.E.(\hat{p}) = sd(\hat{p}) = \sqrt{\frac{pq}{n}}$$

: 표본비율 \hat{p} 은 모비율 p 의 일치추정량

$$P\{|\hat{p} - p| \geq \varepsilon\} \leq \frac{Var(\hat{p})}{\varepsilon^2} = pq/(n\varepsilon^2)$$

: 표본 크기가 큰 경우 $X \sim N(np, npq)$ 표본비율 \hat{p} 는 $N(p, pq/n)$ 근사

추정 - 점 추정

1. 모분산의 추정

: X_1, \dots, X_n 은 모평균 μ , 모분산 σ^2 인 모집단의 확률표본일 때,

$$\begin{aligned} E[\Sigma(X_i - \bar{X})^2] &= E\left[\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right] = E[\Sigma X_i^2 - n\bar{X}^2] \\ &= \Sigma E(X_i^2) - nE(\bar{X}^2) \end{aligned}$$

2. 모분산과 모표준편차 점 추정

: 모분산의 추정량 $= \hat{\sigma}^2 = S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$

: 모표준편차의 추정량 $= \hat{\sigma} = \sqrt{S^2} = S$

추정 - 구간추정

1. 구간추정

: 표본에서 얻어지는 정보를 이용하여 모수가 속할 것으로 기대되는 범위(신뢰구간)를 택하는 과정
: 통계적 추정은 일반적으로 신뢰구간의 추정을 활용

: 모수 θ 에 대하여 $P(a < \theta < b) = 1 - \alpha$ 일 때 구간 (a,b) 을 모수 θ 에 대한 $100(1 - \alpha)\%$ 신뢰구간이라고 한다.

2. 신뢰구간

: 모수를 포함할 것으로 추정한 구간

3. 신뢰수준

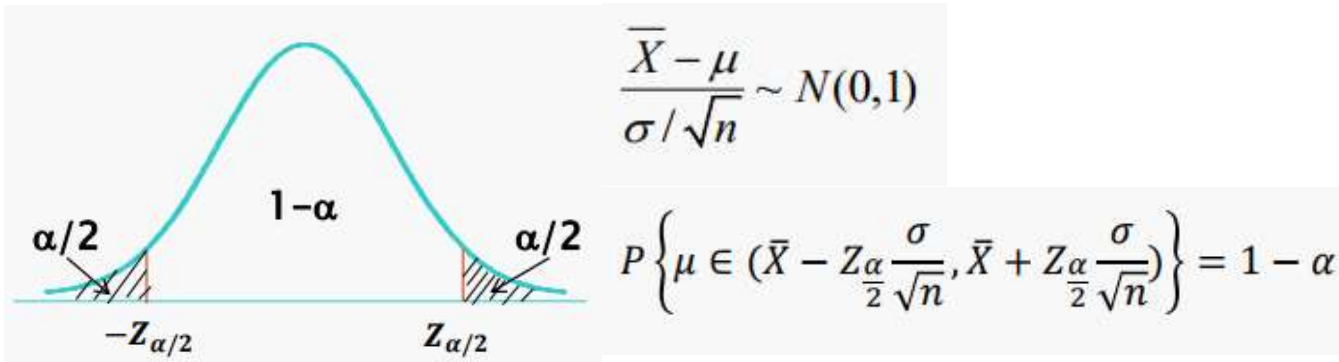
: 신뢰구간이 모수를 포함할 확률 $(1 - \alpha)$ * α : 오차율
: 동일한 표본추출을 통해 구한 신뢰구간들 중 $100 \times (1 - \alpha) \%$ 는 모수를 포함

추정 - 모평균의 구간추정

1. 모분산을 아는 경우

가정) 모분산을 안다.

모집단의 평균이 μ , 분산이 σ^2 인 정규분포
Z통계량을 사용



- 90% 신뢰구간 : $Z_{0.05} = 1.64$
- 95% 신뢰구간 : $Z_{0.025} = 1.96$
- 99% 신뢰구간 : $Z_{0.005} = 2.57$

추정 - 모평균의 구간추정

1. 모분산을 모르는 경우

가정) 모분산을 모른다.

모집단의 평균이 μ , 분산이 σ^2 인 정규분포
t통계량을 사용 (표본 크기가 클 경우 Z통계량을 사용)

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

$$P\left\{-t(n-1, \frac{\alpha}{2}) \leq \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t(n-1, \frac{\alpha}{2})\right\} = 1 - \alpha$$

$$P\left\{\mu \in \left(\bar{X} - t\left(n-1, \frac{\alpha}{2}\right) \frac{S}{\sqrt{n}}, \bar{X} + t\left(n-1, \frac{\alpha}{2}\right) \frac{S}{\sqrt{n}}\right)\right\} = 1 - \alpha$$

추정 - 모평균의 구간추정

1. 모평균의 $100(1-\alpha)\%$ 신뢰구간

1) 표본 크기가 크지 않은 경우

- 모분산 known

$$\left(\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

- 모분산 unknown

$$\left(\bar{X} - t\left(n-1, \frac{\alpha}{2}\right) \frac{S}{\sqrt{n}}, \bar{X} + t\left(n-1, \frac{\alpha}{2}\right) \frac{S}{\sqrt{n}}\right)$$

2) 표본 크기가 큰 경우

- 모분산 known $\left(\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$

- 모분산 unknown $\left(\bar{X} - Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right)$

추정 - 모비율의 구간추정

1. 모비율의 구간추정

: 베르누이분포 $B(1,p)$ 로부터의 크기 n 인 확률표본에 대해,
표본비율 \hat{p} 의 분포는 n 이 클 때, 근사적으로
 $N\left(p, \frac{p(1-p)}{n}\right)$ 를 따름.

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

2. 근사신뢰구간

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

추정 - 모분산의 구간추정

1. 모분산의 구간추정

: X_1, \dots, X_n 이 정규분포 $N(\mu, \sigma^2)$ 로부터의 확률표본,
표본분산에 대하여

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$$P\left\{\chi^2\left(n-1, 1-\frac{\alpha}{2}\right) \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2\left(n-1, \frac{\alpha}{2}\right)\right\} = 1-\alpha$$

2. 정규모집단의 모분산 σ^2 에 대한 $100(1-\alpha)\%$ 신뢰구간

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

추정 - 표본크기 결정

1. 모평균 추정

: 100(1- α)% 오차한계를 d 이하로, 하는데 필요한 표본크기 n

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

$$Z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right) \leq d, \quad n \geq \left(Z_{\frac{\alpha}{2}} \frac{\sigma}{d} \right)^2$$

2. 모비율 추정

: 100(1- α)% 오차한계를 d 이하로, 하는데 필요한 표본크기 n

- p에 대한 사전지식이 없는 경우: $n \geq \frac{1}{4} \left(\frac{Z_{\frac{\alpha}{2}}}{d} \right)^2$

- p에 대한 사전지식(p^*)이 있는 경우: $n \geq p^* q^* \left(\frac{Z_{\frac{\alpha}{2}}}{d} \right)^2$

추정 - 표본크기 결정

1. 최소표본크기 조정

: 모집단 크기가 표본 크기에 비해 상대적으로 작을 경우
최소표본의 크기를 조정

- $100(1-\alpha)\%$ 신뢰구간 가정 하, 최소표본 크기 계산
- 최소표본 크기(=n)와 모집단 크기(=N) 비교
- n/N 이 α 보다 크면 표본크기 조정

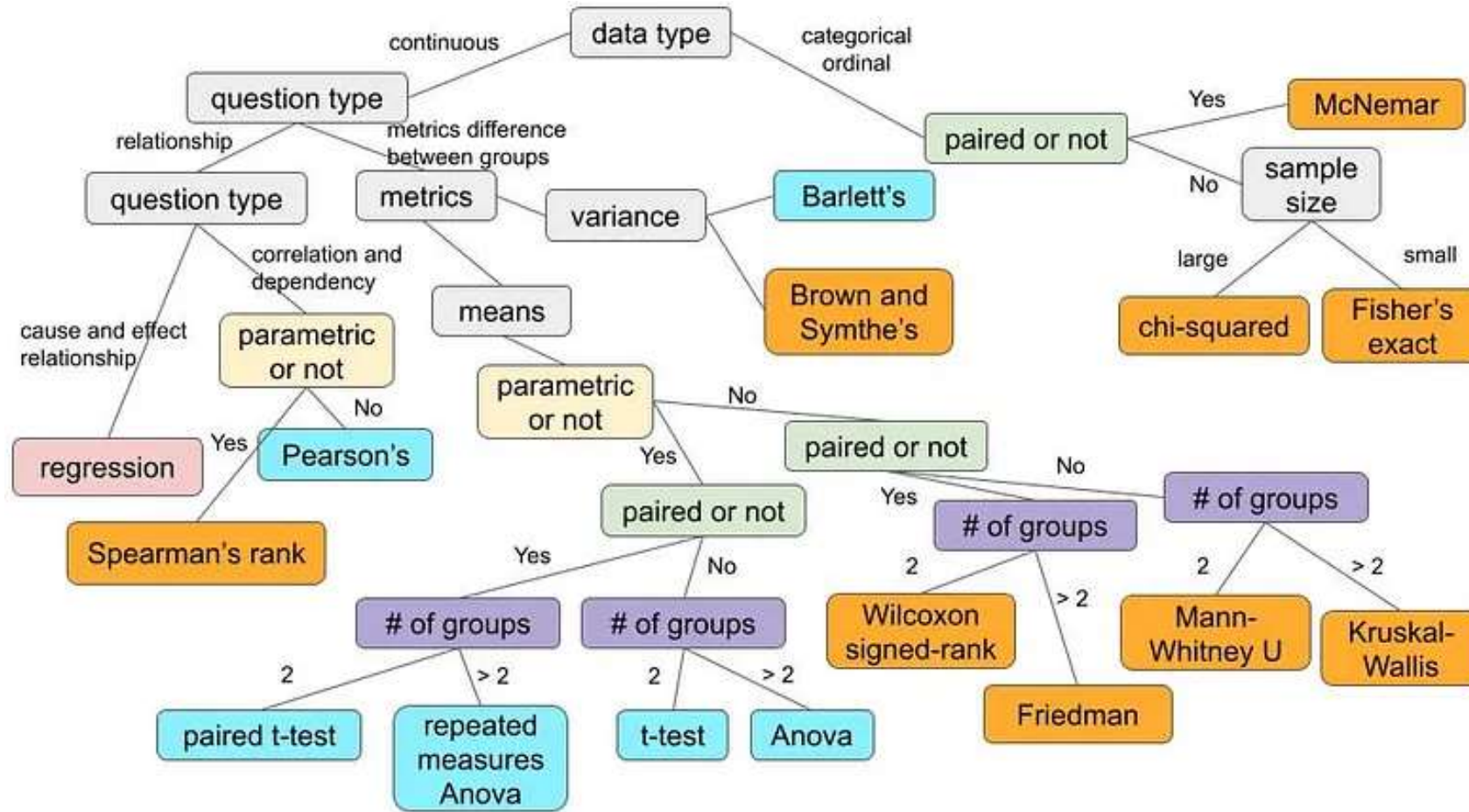
$$n_{adj} = \frac{n}{1 + \frac{n}{N}}$$

1. 통계적 추론

- 신뢰구간: 미지의 모수 추정이 목적
- 가설검정: 모수의 관한 가설의 타당성에 대한 경험적 증거 제시가 목적

2. 가설검정

- : 모수에 관한 귀무/대립가설을 설정한 후 데이터에 따라 어떤 가설이 맞는지 결정하는 통계적 분석
- 비수도권의 출생률이 수도권 출생률보다 낮다
- X팀은 홈구장에서의 승률이 원정 구장에서의 승률보다 높다



Source: <https://medium.com/@rich.tsai1103>

1. 가설의 종류

: 귀무가설 (H_0)

- 검정하는 가설
- 대립가설과 상반되는 가설로, 일반적인 사실을 귀무가설로 설정
- 효과가 없다, 차이가 없다 등의 내용

: 대립가설 (H_1)

- 입증하고자 하는 가설
- 효과가 있다, 차이가 있다 등의 내용

2. 가설의 예시

예) X팀의 홈구장 승률은 60%보다 높은가?

- 귀무가설: X팀의 홈구장 평균 승률은 60%보다 작거나 같다. ($\mu \leq 0.60$)
- 대립가설: X팀의 홈구장 평균 승률은 60%보다 크다. ($\mu > 0.60$)

1. 가설설정 오류

- 제1종 오류 (α)

- : 귀무가설을 채택해야 했음에도 이를 기각할 오류
- : 표본으로부터 얻은 검정결과가 우연에 의해 잘못 판단되었을 가능성
- : α 는 일반적으로 5%로 설정

- 제2종 오류 (β)

- : 귀무가설을 기각해야 했음에도 이를 채택할 오류
- : 실제로는 효과가 없는데 효과가 있다고 잘못 결론 내릴 가능성
- : β 는 일반적으로 10%로 설정

1. 가설설정의 오류

| | 귀무가설 True | 귀무가설 False |
|---------|--------------------|-------------------|
| 귀무가설 채택 | 옳음 ($1-\alpha$) | 제2종 오류(β) |
| 귀무가설 기각 | 제1종 오류(α) | 옳음 ($1-\beta$) |

- 오류를 완벽히 배제할 수는 없음
- 두 가지 오류를 동시에 최소로 할 수는 없음
- 일반적으로 제1종 오류가 더 중요하여 이를 미리 지정한 유의수준 이하로 하는 방식 사용

검정 - 요소

1. 유의수준 (significance level)

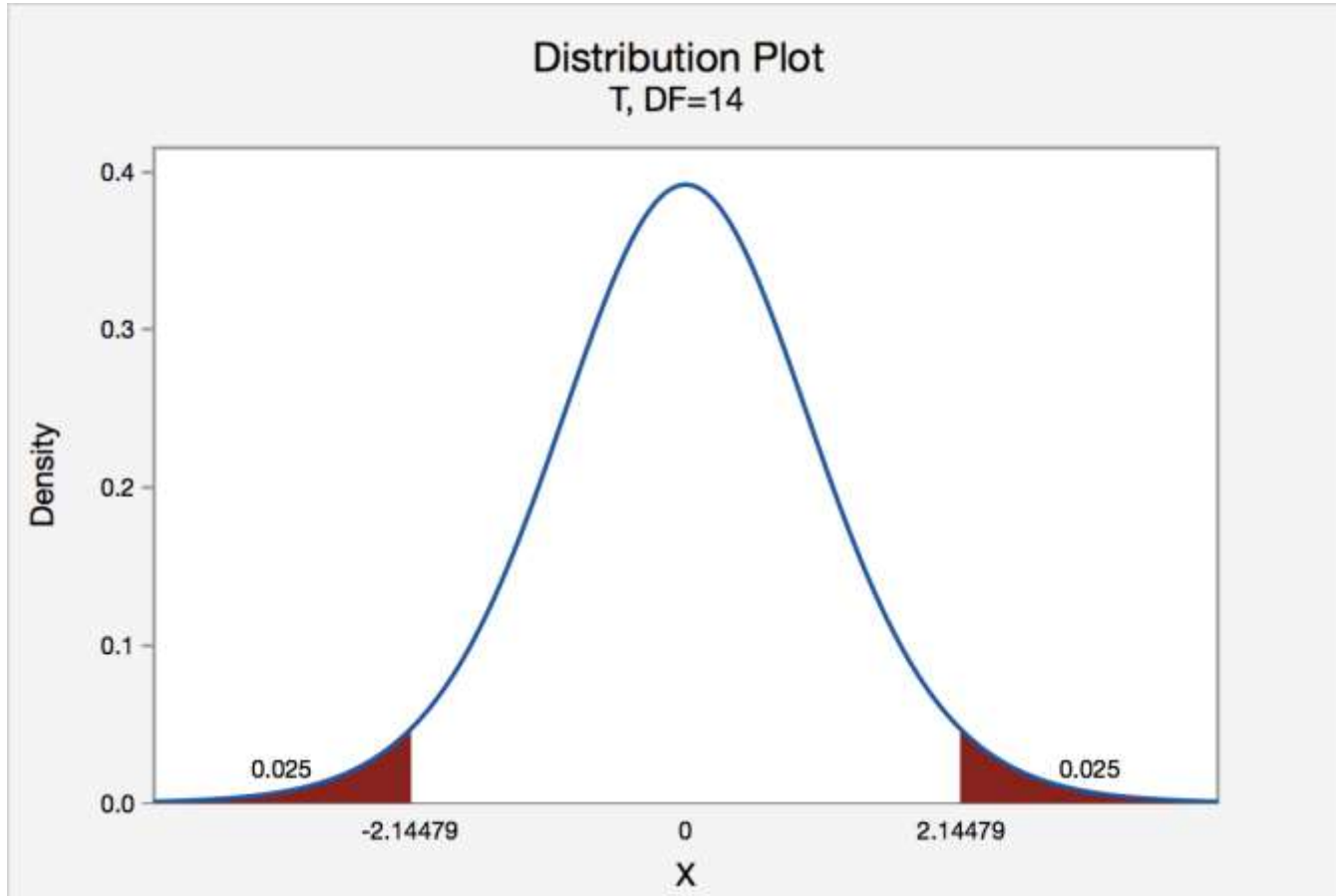
- 제1종 오류를 범할 확률의 최대 허용한계

2. 유의확률 (p-value)

- 귀무가설이 옳다는 가정하에 관측된 사건만큼 혹은 이보다 더 귀무가설에 반하는 사건이 일어날 확률
- 검정통계량 값에 대해 귀무가설을 기각할 수 있는 최소의 유의수준으로 귀무가설이 사실일 확률
- $\alpha > p\text{-value}$: 귀무가설 기각
- $\alpha < p\text{-value}$: 귀무가설 채택

3. 임계값 (critical value)

- 기각역과 채택역을 나누는 경계값
- 기각역 : 귀무가설을 기각하게 되는 검정통계량의 관측값의 영역
- 채택역 : 귀무가설을 채택하게 되는 검정통계량의 관측값의 영역
- 검정통계량의 관측값이 기각역에 속하면 귀무가설 기각



Source: <https://online.stat.psu.edu/statprogram>

검정 - 검정력함수

1. 검정력함수 (power function)

- : 모수의 값에 따른 귀무가설 기각 확률의 변화에 대한 함수
- : 검정통계량이 X , 기각역이 R 이라면

$$\text{검정력함수} = \gamma(\theta) = P\{X \in R | \theta\}$$

- 제1종오류 범할 확률 = $\gamma(\theta)$
- 제2종오류 범할 확률 = $1 - \gamma(\theta)$

: 검정력(power): 귀무가설이 사실이 아닐 때 귀무가설을 기각할 확률

2. 예시) “기각역이 $\bar{X} \leq 10$ 인 검정법의 검정력 함수

$$\gamma(\mu) = P[\bar{X} \leq 10 | \text{true mean} = \mu] = P[Z \leq \frac{10 - \mu}{\frac{\sigma}{\sqrt{n}}}]$$

검정 -절차

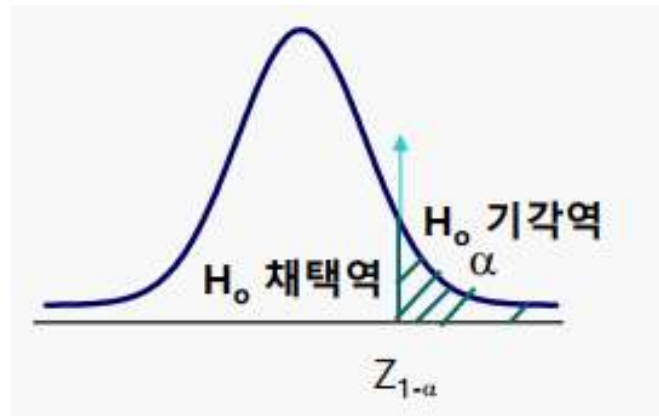
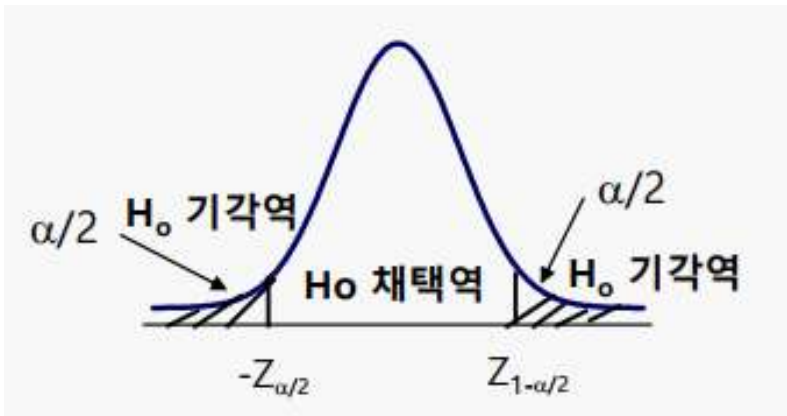
1. 검정할 가설을 설정
2. 유의수준을 설정
3. 임계치를 결정하고 검정통계량과 임계치를 비교
(혹은 유의수준과 유의확률 비교)
4. 유의확률(=p-value)값이 유의수준(α)보다 작으면 귀무가설을 기각

1. 양측검정 (Two-sided)

- 기각역이 각각 왼쪽과 오른쪽 두 부분으로 구성된 가설검정
- 양쪽 기각역의 합 = 유의수준

2. 단측검정 (One-sided)

- 기각역이 한쪽으로만 구성되는 가설검정
- 한쪽 기각역이 유의수준



검정 - 모평균 검정

1. 정규성 가정 만족하는 경우

- 정규분포를 따르는 모집단으로부터 n 개의 표본이 추출되었을 때
- 모분산 known: Z 검정 통계량 사용

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- 모분산 unknown: t 검정 통계량 사용

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

검정 - 모평균 검정

1. 모평균 검정 - 정규 모집단

- 모분산 known: Z 검정 통계량 사용

| 귀무가설 | 대립가설 | 기각역 |
|-----------------------|-----------------------|-------------------------|
| $H_0: \mu \leq \mu_0$ | $H_1: \mu > \mu_0$ | $Z \geq Z_\alpha$ |
| $H_0: \mu \geq \mu_0$ | $H_1: \mu < \mu_0$ | $Z \leq -Z_\alpha$ |
| $H_0: \mu = \mu_0$ | $H_1: \mu \neq \mu_0$ | $ Z \geq Z_{\alpha/2}$ |

- 모분산 unknown: t 검정 통계량 사용

| 귀무가설 | 대립가설 | 기각역 |
|-----------------------|-----------------------|-------------------------------------|
| $H_0: \mu \leq \mu_0$ | $H_1: \mu > \mu_0$ | $T \geq t(n-1, \alpha)$ |
| $H_0: \mu \geq \mu_0$ | $H_1: \mu < \mu_0$ | $T \leq -t(n-1, \alpha)$ |
| $H_0: \mu = \mu_0$ | $H_1: \mu \neq \mu_0$ | $ T \geq t(n-1, \frac{\alpha}{2})$ |

검정 - 모평균 검정

1. 표본의 크기가 큰 임의의 모집단

- 모분산 known: Z 검정 통계량 사용 $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
- 모분산 unknown: Z 검정 통계량 사용 $Z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$

| 귀무가설 | 대립가설 | 기각역 |
|-----------------------|-----------------------|-------------------------|
| $H_0: \mu \leq \mu_0$ | $H_1: \mu > \mu_0$ | $Z \geq Z_\alpha$ |
| $H_0: \mu \geq \mu_0$ | $H_1: \mu < \mu_0$ | $Z \leq -Z_\alpha$ |
| $H_0: \mu = \mu_0$ | $H_1: \mu \neq \mu_0$ | $ Z \geq Z_{\alpha/2}$ |

검정 - 모평균 검정

1. 표본의 크기

- 모분산 known인 정규모집단에서
- $H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$ 혹은 $H_0: \mu \geq \mu_0$, $H_1: \mu < \mu_0$
유의수준 α 인 검정법에서 대립가설 하의 평균값 $\mu = \mu_1$ 이 True라면,
제2종 오류를 범한 확률이 β 이하가 되도록 하려면 표본의 크기 n 은

$$n \geq \left\{ \frac{Z_\alpha + Z_\beta}{\frac{\mu_1 - \mu_0}{\sigma}} \right\}^2$$

검정 - 모평균 검정

1. 예시 - 표본의 크기

- 한 배달업체의 기존 배달시간은 평균 30분, 표준편차는 5분이라고 한다.
새로운 배달방식에 의한 평균 배달시간이 20분이 된다면 큰 개선으로 간주한다.
실제로 $\mu = 20$ 일 때, 잘못 결정하는 확률을 $\beta = 0.1$ 이하로 하고자 할 때,
유의수준 $\alpha = 0.05$ 인 검정법의 표본 크기는?

풀이) $\mu = 20$ 일 때, 검정력을 0.9 이상으로 해야 하므로

$$n \geq \left\{ \frac{Z_{\alpha} + Z_{\beta}}{\frac{\mu_1 - \mu_0}{\sigma}} \right\}^2$$

검정 - 모비율 검정

1. 모비율의 검정

- 표본의 크기가 작은 경우에는 이항검정법을 사용

| 귀무가설 | 대립가설 | 기각역 |
|-------------------|-------------------|--|
| $H_0: p \leq p_0$ | $H_1: p > p_0$ | $X \geq c (P\{X \geq c p = p_0\} \leq \alpha \text{인 } c \text{ 중 최소값})$ |
| $H_0: p \geq p_0$ | $H_1: p < p_0$ | $X \leq c (P\{X \leq c p = p_0\} \leq \alpha \text{인 } c \text{ 중 최대값})$ |
| $H_0: p = p_0$ | $H_1: p \neq p_0$ | $X \leq c_1 \text{ 또는 } X \geq c_2$ 단, $P\{X \leq c_1 p = p_0\} \leq \alpha/2$ 인 c_1 중 최대값 $P\{X \geq c_2 p = p_0\} \leq \alpha/2$ 인 c_2 중 최소값 |

검정 - 모비율 검정

1. 예시 - 모비율의 검정

- 한 질병의 발병률은 3%로 알려져 있다. 시간이 지나 100명을 랜덤추출하여 새롭게 조사를 해보니 감염환자는 2명이다. 이 조사로부터 해당 질병의 발병률이 기존보다 낮아졌는지 유의수준 5%에서 검정

$$H_0: p \geq 0.03, H_1: p < 0.03, \text{ 기각역: } X \leq c$$

포아송 분포를 통해 접근

$$P(X \leq c) \simeq \sum_{x=0}^c \frac{e^{-3} 3^x}{x!} \leq 0.05$$

검정 - 모비율 검정

1. 모비율의 검정

- $X \sim B(n, p)$ 를 따르고, 표본 크기 n 이 충분히 크고 $np_0 \geq 5, nq_0 \geq 5$ 일 때, 모비율에 관한 검정통계량

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

| 귀무가설 | 대립가설 | (근사적으로) 기각역 |
|-------------------|-------------------|-------------------------|
| $H_0: p \leq p_0$ | $H_1: p > p_0$ | $Z > Z_\alpha$ |
| $H_0: p \geq p_0$ | $H_1: p < p_0$ | $Z < -Z_\alpha$ |
| $H_0: p = p_0$ | $H_1: p \neq p_0$ | $ Z \geq Z_{\alpha/2}$ |

검정 - 모분산 검정

1. 모분산의 검정

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

| 귀무가설 | 대립가설 | 기각역 |
|---------------------------------|---------------------------------|--|
| $H_0: \sigma^2 \leq \sigma^2_0$ | $H_1: \sigma^2 > \sigma^2_0$ | $\chi^2 \geq \chi^2(n-1, \alpha)$ |
| $H_0: \sigma^2 \geq \sigma^2_0$ | $H_1: \sigma^2 < \sigma^2_0$ | $\chi^2 \leq \chi^2(n-1, 1-\alpha)$ |
| $H_0: \sigma^2 = \sigma^2_0$ | $H_1: \sigma^2 \neq \sigma^2_0$ | $\chi^2 \geq \chi^2\left(n-1, \frac{\alpha}{2}\right)$ 또는 $\chi^2 \leq \chi^2\left(n-1, 1-\frac{\alpha}{2}\right)$ |

검정 - 모분산 검정

1. 예시 - 모분산의 검정

- 어느 측정기의 분산이 0.1이고, 새 측정기의 분산이 0.05일 때,
새 측정기의 분산이 0.1보다 작다고 할 수 있는지 유의수준 5%에서
검정하라 (새 측정기의 측정 데이터는 총 20개)

$$H_0: \sigma^2 = 0.1, \quad H_1: \sigma^2 < 0.1$$

$$\text{검정통계량: } \chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(20-1)(0.05)}{0.1} = 9.5$$

$$\text{기각역: } \chi^2 \leq \chi_{0.95}^2 (20-1) = 10.117...$$

두 모집단 비교

1. 두 집단의 비교

- 모평균 비교, 모비율 비교, 모분산 비교 등
- 두 집단의 비교에는 분산이 고려되어야 함
 - 분산이 고려되어야 보다 객관적인 비교가 가능

2. 두 모평균의 비교

- 두 모평균 비교
 - 모분산 Known & 정규모집단
 - 모분산 Unknown & 정규모집단
 - 모분산 같음
 - 모분산 다름
- 짝을 이룬 표본 비교

두 모집단 비교 - 모평균

1. 두 모평균 차이 추론: 모분산 known

X_1, \dots, X_n 과 Y_1, \dots, Y_n 이 각각 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)$ 을 따르고 서로 독립

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$$

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- 우리의 관심 대상은 두 모평균의 차이 ($= \mu_1 - \mu_2$)
- 추정량: $\bar{X} - \bar{Y}$

$$\text{- 검정통계량: } Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

두 모집단 비교 - 모평균

1. 예시) 두 모평균 차이 추론: 모분산 known

- 그룹 A와 그룹B의 키 평균이 같은가?

($\bar{x} = 170$, $\bar{y} = 173$, $n_1 = 10$, $n_2 = 10$, $\sigma_1^2 = 3$, $\sigma_2^2 = 10$)

- 가설

- 검정통계량:
$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \doteq 2.63 > 1.96$$

- 신뢰구간

두 모집단 비교 - 모평균

1. 두 모평균 차이 추론: 소표본, 모분산 unknown, 등분산

X_1, \dots, X_n 과 Y_1, \dots, Y_n 이 각각 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)$ 을 따르고 서로 독립

$$S_1^2 = \frac{\sum_i^{n_1} (X_i - \bar{X})^2}{(n_1 - 1)}, \quad S_2^2 = \frac{\sum_i^{n_2} (Y_i - \bar{Y})^2}{(n_2 - 1)}$$

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

이를, 합동분산으로 정리하면,

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

등분산 σ^2 합동분산의 추정치 $S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

두 모집단 비교 - 모평균

1. 두 모평균 차이 추론: 소표본, 모분산 unknown, 등분산

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{1/n_1 + 1/n_2}} \sim N(0, 1) \quad \chi^2 = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

$Z \sim N(0, 1)$, $V \sim \chi^2(k)$ Z 와 V 는 서로 독립

$$T = \frac{Z}{\sqrt{\frac{V}{k}}} \sim t(k)$$

- 검정통계량

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$$

두 모집단 비교 - 모평균

1. 예시) 두 모평균 차이 추론: 소표본, 모분산 unknown, 등분산

- 그룹 A와 그룹B의 키 평균이 같은가?
($\bar{x} - \bar{y} = 5$, $n_1 = 15$, $n_2 = 10$, $S_p = 3$)
- 가설
- 검정통계량:
$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \doteq 4.08 > 2.069$$
- 신뢰구간

두 모집단 비교 - 모평균

1. 두 모평균 차이 추론: 소표본, 모분산 unknown, 이분산

- 검정통계량

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

근사적으로 t분포

자유도: Satterwaite 자유도

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1 - 1} + \left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2 - 1}\right)}$$

두 모집단 비교 - 모평균

1. 두 모평균 차이 추론: 표본 크기가 크다면

- 모평균의 차($\mu_1 - \mu_2$)의 $100(1-\alpha)\%$ 신뢰구간

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{S_1^2/n_1 + S_2^2/n_2}$$

- 귀무가설 $H_0: \mu_1 - \mu_2 = \delta$ 을 검정하기 위한 검정통계량

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

두 모집단 비교 - 대응비교

1. 짝을 이룬 표본의 차이

- 같은 개체에 대해 실험 전/후 측정 값의 차이를 추론
- 두 집단 독립 X
- **대응표본**: 서로 독립 X 비슷한 성질의 표본
- **대응비교**: 대응표본을 사용하여 두 모집단의 평균을 비교

2. 모평균의 차 ($\mu_1 - \mu_2 = \delta$)에 관한 추론

- 대응비교 $D_i = X_i - Y_i$, 서로 독립이고 $N(\delta, \sigma_D^2)$ 가정
- $\delta = \mu_1 - \mu_2$ 에 대한 $100(1 - \alpha)\%$ 신뢰구간

$$\bar{D} \pm t_{\frac{\alpha}{2}}(n - 1)S_D/\sqrt{n}$$

$$\bar{D} = \frac{1}{n}\sum D_i, \quad S_D^2 = \frac{\sum (D_i - \bar{D})^2}{n - 1}$$

- 귀무가설 $H_0: \mu_1 - \mu_2$ 를
검정하기 위한 검정통계량 $T = \frac{\bar{D} - \delta}{S_D/\sqrt{n}} \sim t(n - 1)$

두 모집단 비교 - 두 모비율 비교

1. 두 모집단의 비율 비교

- 두 독립 표본으로부터 비율 \hat{p}_1 과 \hat{p}_2
- 표본비율의 평균과 분산

$$E(\hat{p}_1) = p_1, \quad E(\hat{p}_2) = p_2$$

$$Var(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1}, \quad Var(\hat{p}_2) = \frac{p_2(1-p_2)}{n_2},$$

2. 표본의 크기가 큰 경우

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \approx N(0, 1)$$

두 모집단 비교 - 두 모비율 비교

1. $p_1 - p_2$ 신뢰구간 (표본 크기 큰 경우)

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_1(1 - \widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1 - \widehat{p}_2)}{n_2}}$$

2. 표본의 크기가 큰 경우

$$H_0: p_1 = p_2 (= p)$$

H_0 하에서

$$E(\widehat{p}_1 - \widehat{p}_2) = 0, \quad Var(\widehat{p}_1 - \widehat{p}_2) = p(1 - p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

공통 모비율 p 의 합동추정량 $\hat{p} = \frac{X + Y}{n_1 + n_2}$

$$H_0: p_1 = p_2 \text{ 검정통계량 } Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1 - \widehat{p})} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

4 두 모집단 비교 - 두 모비율 비교

1. 예시) 두 모집단 비율의 검정

| | 강의A | 강의B |
|-----|-----|-----|
| 합격 | 70 | 90 |
| 불합격 | 30 | 60 |
| 합계 | 100 | 150 |

- 합격률 차에 대한 95% 신뢰구간

$$\widehat{p}_1 = 0.7, \quad \widehat{p}_2 = 0.6$$

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

두 모집단 비교 - 두 모비율 비교

1. 예시) 두 모집단 비율의 검정

- 유의수준 5% 하에서 강의A를 수강한 학생의 합격률이 강의B를 수강한 학생의 합격률보다 높은가?

$$H_0: p_1 = p_2, \quad H_1: p_1 > p_2$$

귀무가설 하에서의 공통 모비율 p 의 합동추정량

$$\hat{p} = \frac{70 + 90}{100 + 150} = 0.64$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.1}{\sqrt{0.64 * 0.36} \sqrt{\frac{1}{100} + \frac{1}{150}}} = 1.6137$$

두 모집단 비교 - 두 모분산 비교

1. 두 모집단의 분산 비교

- 두 집단의 분산의 차이는 두 집단의 평균 차이 검정에 영향을 줌
- 분산의 동일성 검정
 - 여러 모집단에 대한 분산의 동일성 여부를 검정
 - 분산의 동일성 검정은 ANOVA분석의 기본 가정
 - 등분산 가정이 성립해야 유의미한 분석 가능

2. 두 모분산 비교의 가정

- X_1, \dots, X_n 과 Y_1, \dots, Y_n 이 각각 $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)$ 을 따르고 서로 독립 (정규분포 아닐 경우 Levene's Test 사용)

3. 귀무가설

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \Rightarrow F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \sim F(n_1 - 1, n_2 - 1)$$

두 모집단 비교 - 두 모분산 비교

1. 두 모집단의 분산 비교

검정통계량 : $F = \frac{S_1^2}{S_2^2}$

가설검정

$$H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1 \quad F \geq F(n_1 - 1, n_2 - 1; \alpha): \text{기각}$$

$$H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1 \quad F \leq F(n_1 - 1, n_2 - 1; 1 - \alpha): \text{기각}$$

$$H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \quad F \geq F\left(n_1 - 1, n_2 - 1; \frac{\alpha}{2}\right) \text{ or } \\ F \leq F\left(n_1 - 1, n_2 - 1; 1 - \frac{\alpha}{2}\right): \text{기각}$$

신뢰구간

$$\frac{S_1^2}{S_2^2} \frac{1}{F(n_1 - 1, n_2 - 1; \alpha/2)} \leq \sigma_1^2 / \sigma_2^2 \leq \frac{S_1^2}{S_2^2} F(n_2 - 1, n_1 - 1; \alpha/2)$$

4 두 모집단 비교 - 두 모분산 비교

1. 두 모집단의 분산 비교

신뢰구간

$$\frac{S_1^2}{S_2^2} \frac{1}{F(n_1 - 1, n_2 - 1; \alpha/2)} \leq \sigma_1^2 / \sigma_2^2 \leq \frac{S_1^2}{S_2^2} F(n_2 - 1, n_1 - 1; \alpha/2)$$

증명)

$$X_1, \dots, X_n \sim N(\mu_1, \sigma_1^2), Y_1, \dots, Y_n \sim N(\mu_2, \sigma_2^2) \Rightarrow \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi_{n_1-1}^2, \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi_{n_2-1}^2$$

$$F = \frac{\frac{(n_2 - 1)S_2^2}{\sigma_2^2}}{\frac{(n_1 - 1)S_1^2}{\sigma_1^2}} = \frac{\sigma_1^2 S_2^2}{\sigma_2^2 S_1^2} \sim F(n_2 - 1, n_1 - 1) \Rightarrow P \left[F_{1-\frac{\alpha}{2}}(n_2 - 1, n_1 - 1) \leq \frac{\sigma_1^2 S_2^2}{\sigma_2^2 S_1^2} \leq F_{\frac{\alpha}{2}}(n_2 - 1, n_1 - 1) \right] = 1 - \alpha$$

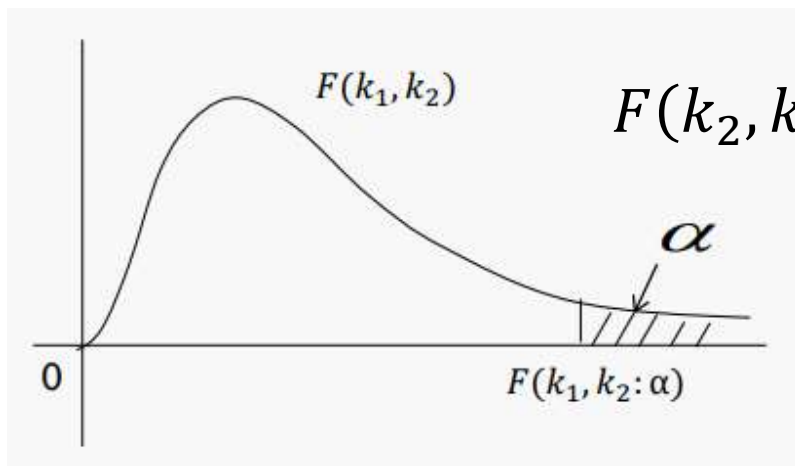
이 때, 양변을 $\frac{S_2^2}{S_1^2}$ 로 나눠주고, $F_{1-\frac{\alpha}{2}}(n_2 - 1, n_1 - 1) = \frac{1}{F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)}$ 을 활용하면 위와 같은 신뢰구간이 나온다.

두 모집단 비교 - 두 모분산 비교

1. F분포

- : F-분포는 두 정규모집단의 분산을 비교하는 추론에 사용
- : V_1 과 V_2 는 각각 자유도 k_1, k_2 인 카이제곱분포를 따르는 독립인 확률변수

$$F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2) \quad \frac{1}{F} = \frac{V_2/k_2}{V_1/k_1} \sim F(k_2, k_1)$$



$$F(k_2, k_1; 1 - \alpha) = \frac{1}{F(k_1, k_2; \alpha)}$$

두 모집단 비교 - 두 모분산 비교

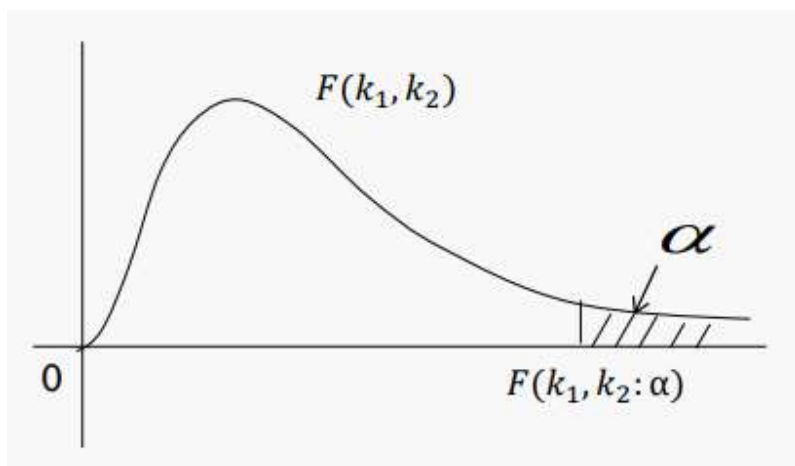
1. 예시) 두 모집단 분산 차이 가설검정 (단측)

$$: n_A = 21, S_A^2 = 10, \quad n_B = 25, S_B^2 = 15$$

$$H_0: \sigma_A^2 = \sigma_B^2, \quad H_1: \sigma_A^2 > \sigma_B^2$$

$$F = \frac{S_A^2}{S_B^2} = \frac{10}{15} = 0.666$$

$$F(n_B - 1, n_A - 1, \alpha) = F(24, 20, 0.05) = 2.08$$



두 모집단 비교 - 두 모분산 비교

1. 예시) 두 모집단 분산 차이 가설검정 (단측)

$$: n_A = 21, S_A^2 = 10, \quad n_B = 25, S_B^2 = 15$$

$$H_0: \sigma_A^2 = \sigma_B^2, \quad H_1: \sigma_A^2 < \sigma_B^2$$

$$F = \frac{S_A^2}{S_B^2} = \frac{10}{15} = 0.666$$

$$\frac{1}{F(n_A-1, n_B-1, \alpha)} = \frac{1}{F(20, 24, 0.05)} = F(24, 20, 0.95) = \frac{1}{2.03}$$

분산 분석

1. 분산분석 (Analysis of Variance)

- 독립변수의 수준(범주)으로 나뉜 집단 간 평균 차이를 검정
- 특성값의 산포를 인자별로 분해하여 어느 인자가 큰 영향을 주는지
- 반응변수
 - 처리(treatment)에 의해 변화하는, 연구대상이 되는 변수
- 인자 (=요인, factor)
 - 반응변수에 영향을 주는 변수
 - 독립변수, 설명변수
- 처리(treatment)
 - 요인의 특정 값 (특정 실험 조건)
 - 수준(level)

2. 기본 가정

- 각 집단의 모집단 분포는 정규분포
- 각 집단의 모집단의 분산 같음
- 각 모집단 내의 오차와 모집단 간 오차는 독립

1. 분산분석 (Analysis of Variance)

- 전체 변동 = 그룹간 변동 + 그룹내 변동
 - 그룹간 변동 = 요인에 의한 효과
 - 그룹내 변동 = 오차에 의한 효과
- 그룹간 분산과 그룹내 분산의 비교

2. 종류

- 일원배치분산분석 (one-way anova)
 - 2 sample t test와 같은 목적이지만 anova는 2개 이상의 모집단의 평균 비교
 - 인자가 하나인 경우, 두 개 이상의 모집단 평균이 서로 동일한지 검정
- 이원배치분산분석 (two-way anova)
 - 인자가 두개인 경우에 사용
 - 인자의 수준 및 교호 작용에 의한 영향 파악 가능

분산 분석 - 일원배치

1. 일원배치 분산분석

- 한 개의 요인
- 가정
 - k개의 모집단은 독립 & 정규분포
 - 각 집단의 모집단 평균은 서로 다를 수 있으나 분산은 같음

2. 가설검정

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

H_1 : 적어도 하나의 평균은 다르다

분산 분석 - 일원배치

1. 일원배치 분산분석 - 반복수가 같은 경우

| | 처리 1 | 처리 2 | ... | 처리 k |
|------|--|-------------|-----|-------------|
| 데이터 | y_{11} | y_{21} | ... | y_{k1} |
| | y_{12} | y_{22} | ... | y_{k2} |
| | ... | ... | ... | ... |
| | y_{1n} | y_{2n} | ... | y_{kn} |
| 크기 | n | n | ... | n |
| 평균 | \bar{y}_1 | \bar{y}_2 | ... | \bar{y}_k |
| 표준편차 | s_1 | s_2 | ... | s_k |
| 전체 | $\bar{y} = (y_{11} + y_{12} + \dots + y_{kn})/N$ | | | |

$$Y_{ij}: \mu_i + \epsilon_{ij}, \quad (j = 1, 2, \dots, n), \quad \epsilon \sim N(0, \sigma^2)$$

분산 분석 - 일원배치

1. 관찰값 모형

$$Y_{ij}: \mu_i + \epsilon_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

μ : 전체 모평균

μ_i : i번째 처리의 모평균

α_i : 처리 i의 효과, 처리 i에서의 모평균 μ_i 가 전체 모평균 μ 로부터 어느정도 치우쳤는지

$$- \sum \alpha_i = 0$$

ϵ_{ij} : i번째 처리의 j번째 반응값이 가치는 오차,
상호 독립, $\epsilon_{ij} \sim N(0, \sigma^2)$

분산 분석 - 일원배치

1. 관찰값 모형

$$Y_{ij}: \mu_i + \epsilon_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

오차항의 가정

- 1) 정규성: 정규분포
- 2) 독립성: 오차항 간 독립
- 3) 비편향성: 기댓값은 0 (Bias 없음)
- 4) 등분산성: 모든 오차항의 분산 동일

2. 귀무가설 수립

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$\rightarrow H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k$$

분산 분석 - 일원배치

1. 제곱합 분해

$$y_{ij}: \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

- 1) $(\bar{y}_i - \bar{y})$: i번째 집단의 평균과 전체평균 간의 차이 = 처리효과의 크기
- 2) $(y_{ij} - \bar{y}_i)$: 각 관찰값과 각 집단평균 간의 차이 = 잔차

$$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

$$\Rightarrow \sum \sum (y_{ij} - \bar{y})^2 = \sum \sum (\bar{y}_i - \bar{y})^2 + \sum \sum (y_{ij} - \bar{y}_i)^2$$

- 제곱합: **SST = SS_t + SS_e**
 - (총 분산) = (처리간 분산) + (처리내 분산)
- 자유도: $kn-1 = (k-1) + k(n-1)$

1. 제곱합 분해 - 증명

$$\begin{aligned}\sum\sum(y_{ij} - \bar{y})^2 &= \sum\sum \left((\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i) \right)^2 \\ &= \sum\sum (\bar{y}_i - \bar{y})^2 + 2\sum\sum (\bar{y}_i - \bar{y})(y_{ij} - \bar{y}_i) + \sum\sum (y_{ij} - \bar{y}_i)^2\end{aligned}$$

분산 분석 - 일원배치

1. 평균 제곱

- 제곱합을 자유도로 나눈 값 (분산의 추정량)
- MSt (처리평균제곱)

$$MSt = \frac{SSt}{k-1} = \frac{n\sum(\bar{y}_i - \bar{y})^2}{k-1}$$

- MSE (잔차평균제곱)

$$MSt = \frac{SSt}{k-1} = \frac{n\sum(\bar{y}_i - \bar{y})^2}{k-1}$$

- F-통계량

$$F_{k-1, k(n-1)} = \frac{MSt}{MSE}$$

분산 분석 - 일원배치

1. ANOVA Table

| | 제곱합 | 자유도 | 평균제곱합 | F |
|----|---------|--------|--------------|---------|
| 처리 | SSt | K-1 | SSt/(K-1) | MSt/MSE |
| 잔차 | SSE | K(n-1) | SSE/(K(n-1)) | |
| 총 | SSt+SSE | Kn-1 | | |

- 각 처리 집단의 평균이 같다면 SSt는 작아짐 (SSE 커짐)
- 각 처리 집단의 평균이 다르다면 SSt는 커짐 (SSE 작아짐)

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k$$

- 검정통계량 $F = \frac{MSt}{MSE}$
- $F > F_\alpha(k-1, k(n-1)) \rightarrow$ 귀무가설 기각

분산 분석 - 일원배치

1. 독립변수의 설명력

- 상관계수 $\eta^2 = \frac{SS_t}{SST}$
- 독립변수의 설명력을 나타냄

2. 사후검정

- 어떤 집단 간에서 유의한 차이가 발생했는지 사후적으로 분석
- Post hoc comparison
- LSD, TUKEY, DUNCAN, 등

분산 분석 - 일원배치

1. 예시) $\alpha = 0.05$

| 그룹1 | 그룹2 | 그룹3 | | 제곱합 | 자유도 | 평균제곱합 | F |
|-----|-----|-----|----|--------|-----|-------|-------|
| 85 | 91 | 79 | 처리 | 192.2 | 2 | 96.1 | 2.358 |
| 86 | 92 | 78 | | | | | |
| 88 | 93 | 88 | 잔차 | 1100.6 | 27 | 40.8 | |
| 75 | 85 | 94 | | | | | |
| 78 | 87 | 92 | 총계 | 1292.8 | 29 | 44.58 | |
| 94 | 84 | 85 | | | | | |
| 98 | 82 | 83 | | | | | |
| 79 | 88 | 85 | | | | | |
| 71 | 95 | 82 | | | | | |
| 80 | 96 | 81 | | | | | |

독립변수 설명력: $\frac{192.2}{1292.8} = 0.1486$

분산 분석 - 일원배치

1. 일원배치 분산분석 - 반복수가 다른 경우

| | 처리 1 | 처리 2 | ... | 처리 k |
|------|--|-------------|-----|-------------|
| 데이터 | y_{11} | y_{21} | ... | y_{k1} |
| | y_{12} | y_{22} | ... | y_{k2} |
| | ... | ... | ... | ... |
| | y_{1n_1} | y_{2n_2} | ... | y_{kn_k} |
| 크기 | n_1 | n_2 | ... | n_k |
| 평균 | \bar{y}_1 | \bar{y}_2 | ... | \bar{y}_k |
| 표준편차 | s_1 | s_2 | ... | s_k |
| 전체 | $n = (n_1 + n_2 + \dots + n_k)$ $\bar{y} = (y_{11} + y_{12} + \dots + y_{kn_k})/n$ | | | |

$$Y_{ij} : \mu_i + \epsilon_{ij}, \quad (j = 1, 2, \dots, n), \quad \epsilon \sim N(0, \sigma^2)$$

분산 분석 - 일원배치

1. 관찰값 모형 - 반복 수 다름

$$Y_{ij}: \mu_i + \epsilon_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$(i = 1, 2, \dots, k; j = 1, 2, \dots, n_i)$$

μ : 전체 모평균

μ_i : i 번째 처리의 모평균

α_i : 처리 i 의 효과, 처리 i 에서의 모평균 μ_i 가 전체 모평균 μ 로부터 어느정도 치우쳤는지

$$- \sum_{i=1}^k n_i \alpha_i = 0$$

ϵ_{ij} : i 번째 처리의 j 번째 반응값이 가치는 오차,
상호 독립, $\epsilon_{ij} \sim N(0, \sigma^2)$

분산 분석 - 일원배치

1. 제곱합 분해

$$y_{ij}: \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i), (i = 1, 2, \dots, k; j = 1, 2, \dots, n_i)$$

- 1) $(\bar{y}_i - \bar{y})$: i번째 집단의 평균과 전체평균 간의 차이 = 처리효과의 크기
- 2) $(y_{ij} - \bar{y}_i)$: 각 관찰값과 각 집단평균 간의 차이 = 잔차

$$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

$$\Rightarrow \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- 제곱합: **SST = SSt + SSE**
 - (총 분산) = (처리간 분산) + (처리내 분산)
- 자유도: $kn-1 = (k-1) + k(n-1)$

분산 분석 - 일원배치

1. ANOVA Table

| | 제곱합 | 자유도 | 평균제곱합 | F |
|----|-----------|------------------------|------------------------------|-------------|
| 처리 | SSt | K-1 | SSt/(K-1) | MSt/ MSE |
| 잔차 | SSE | $\sum_{i=1}^k n_i - k$ | $SSE/(\sum_{i=1}^k n_i - k)$ | |
| 총 | SSt + SSE | $\sum_{i=1}^k n_i$ | | |

- 각 처리 집단의 평균이 같다면 SSt는 작아짐 (SSE 커짐)
- 각 처리 집단의 평균이 다르다면 SSt는 커짐 (SSE 작아짐)

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k$$

- 검정통계량 $F = \frac{MSt}{MSE}$
- $F > F_{\alpha}(k-1, \sum_{i=1}^k n_i - k) \rightarrow$ 귀무가설 기각

분산 분석 - 이원배치 (반복X)

1. 이원배치 분산분석

- 두 개의 요인 (A, B)
 - 요인 A(수준 p개)
 - 요인 B(수준 q개)

2. 가설검정

$$H_0: \mu_{A1} = \mu_{A2} = \dots = \mu_{Ap}$$

$$\mu_{B1} = \mu_{B2} = \dots = \mu_{Bq}$$

H_1 : 적어도 요인 A 하나의 평균은 다르다
적어도 요인 B 하나의 평균은 다르다

분산 분석 - 이원배치 (반복X)

| | B_1 | ... | B_j | ... | B_q | 평균 |
|-------|----------------|-----|----------------|-----|----------------|----------------|
| A_1 | y_{11} | ... | y_{1j} | ... | y_{1q} | $\bar{y}_{1.}$ |
| ... | | | | | | ... |
| A_i | y_{i1} | ... | y_{ij} | ... | y_{iq} | $\bar{y}_{i.}$ |
| ... | | | | | | ... |
| A_p | y_{p1} | ... | y_{pj} | ... | y_{pq} | $\bar{y}_{p.}$ |
| 평균 | $\bar{y}_{.1}$ | | $\bar{y}_{.j}$ | | $\bar{y}_{.q}$ | $\bar{y}_{..}$ |

$$\bar{y}_{i.} = \frac{1}{q} \sum_{j=1}^q y_{ij} \quad (i = 1, 2, \dots, p) \quad \bar{y}_{.j} = \frac{1}{p} \sum_{i=1}^p y_{ij} \quad (j = 1, 2, \dots, q)$$

$$\bar{y}_{..} = \frac{1}{pq} \sum_{i=1}^p \sum_{j=1}^q y_{ij}$$

분산 분석 - 이원배치 (반복X)

1. 관찰값 모형

$$Y_{ij}: \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

μ : 전체 모평균

α_i : A의 i번째 수준 효과

$$- \sum \alpha_i = 0$$

β_j : B의 j번째 수준 효과

$$- \sum \beta_j = 0$$

ϵ_{ij} : 오차항

상호 독립, $\epsilon_{ij} \sim N(0, \sigma^2)$

2. 귀무가설 수립

$$H_0: \mu_{A1} = \mu_{A2} = \cdots = \mu_{Ap} \Rightarrow H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_p$$

$$\mu_{B1} = \mu_{B2} = \cdots = \mu_{Bq} \Rightarrow H_0: \beta_1 = \beta_2 = \cdots = \beta_q$$

분산 분석 - 이원배치 (반복X)

1. 제곱합 분해

$$(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

$$SST = \sum_i^p \sum_j^q (y_{ij} - \bar{y}_{..})^2$$

$$SS_A = q \sum_i^p (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SS_B = p \sum_j^q (\bar{y}_{.j} - \bar{y}_{..})^2$$

$$SSE = \sum_i^p \sum_j^q (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

분산 분석 - 이원배치 (반복X)

1. ANOVA Table

| | 제곱합 | 자유도 | 평균제곱합 | F |
|----|--------|--------------|--------------------------|--------------|
| A | SS_A | $p-1$ | $MS_A = SS_A/(p-1)$ | MS_A / MSE |
| B | SS_B | $q-1$ | $MS_B = SS_B/(q-1)$ | MS_B / MSE |
| 잔차 | SSE | $(p-1)(q-1)$ | $MSE = SSE/[(p-1)(q-1)]$ | |
| 총 | SST | $pq-1$ | | |

2. 가설검정

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p$$

- 검정통계량 $F = \frac{MS_A}{MSE}$
- $F > F_\alpha(p-1, (p-1)(q-1)) \rightarrow$ 귀무가설 기각

분산 분석 - 이원배치 (반복O)

1. 이원배치 분산분석

- 두 개의 요인 (A, B)
 - 요인 A(수준 p개)
 - 요인 B(수준 q개)
- 주 효과 분석
 - 독립변수 개별 효과 분석
- 상호작용 효과
 - 주 효과들 간 교호작용 분석

분산 분석 - 이원배치 (반복O)

| | B_1 | B_2 | ... | B_q | 전체 |
|-------|-------------------------------|-------------------------------|-----|-------------------------------|-----------------|
| A_1 | y_{111} ... y_{11n} | y_{121} ... y_{12n} | ... | y_{1q1} ... y_{1qn} | |
| 평균 | $\bar{y}_{11.}$ | $\bar{y}_{12.}$ | ... | $\bar{y}_{1q.}$ | $\bar{y}_{1..}$ |
| A_2 | y_{211} ... y_{21n} | y_{221} ... y_{22n} | ... | y_{2q1} ... y_{2qn} | |
| 평균 | $\bar{y}_{21.}$ | $\bar{y}_{22.}$ | ... | $\bar{y}_{2q.}$ | $\bar{y}_{2..}$ |
| | | | | | |
| A_p | y_{p11} ... y_{p1n} | y_{p21} ... y_{p2n} | ... | y_{pq1} ... y_{pqn} | |
| 평균 | $\bar{y}_{p1.}$ | $\bar{y}_{p2.}$ | ... | $\bar{y}_{pq.}$ | $\bar{y}_{p..}$ |
| 전체 | $\bar{y}_{.1.}$ | $\bar{y}_{.2.}$ | ... | $\bar{y}_{.q.}$ | $\bar{y}_{...}$ |

분산 분석 - 이원배치 (반복O)

1. 관찰값 모형

$$Y_{ijk}: \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}$$

$$(i = 1, 2, \dots, p; j = 1, 2, \dots, q; k = 1, 2, \dots, n)$$

μ : 전체 모평균

α_i : A의 i번째 수준 효과

$$- \sum \alpha_i = 0$$

β_j : B의 j번째 수준 효과

$$- \sum \beta_j = 0$$

γ_{ij} : A(i)와 B(j)의 교호작용 효과

$$- \sum_{i=1}^p \gamma_{ij} = \sum_{j=1}^q \gamma_{ij} = 0$$

ϵ_{ij} : 오차항

상호 독립, $\epsilon_{ij} \sim N(0, \sigma^2)$

분산 분석 - 이원배치 (반복O)

1. 가설 수립

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p$$

$$H_0: \beta_1 = \beta_2 = \dots = \beta_q$$

$$H_0: \gamma_{ij} = 0$$

2. 제곱합 분해

$$(y_{ijk} - \bar{y}_{...}) = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

$$SST = SS_A + SS_B + SS_{A \times B} + SSE$$

분산 분석 - 이원배치 (반복O)

1. 제곱합 분해

$$(y_{ijk} - \bar{y}_{...}) = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

$$SST = SS_A + SS_B + SS_{A \times B} + SSE$$

$$SST = \sum_i^p \sum_j^q \sum_k^n (y_{ijk} - \bar{y}_{...})^2$$

$$SS_A = qn \sum_i^p (\bar{y}_{i..} - \bar{y}_{...})^2$$

$$SS_B = pn \sum_j^q (\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SS_{A,B} = n \sum_i^p \sum_j^q (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$SSE = \sum_i^p \sum_j^q \sum_k^n (y_{ijk} - \bar{y}_{ij.})^2$$

분산 분석 - 이원배치 (반복O)

1. ANOVA Table

| | 제곱합 | 자유도 | 평균제곱합 | F |
|----|------------|--------------|------------------------------------|------------------|
| A | SS_A | $p-1$ | $MS_A = SS_A/(p-1)$ | MS_A / MSE |
| B | SS_B | $q-1$ | $MS_B = SS_B/(q-1)$ | MS_B / MSE |
| 교호 | $SS_{A,B}$ | $(p-1)(q-1)$ | $MS_{A,B} = SS_{A,B}/[(p-1)(q-1)]$ | $MS_{A,B} / MSE$ |
| 잔차 | SSE | $pq(n-1)$ | $MSE = SSE/[pq(n-1)]$ | |
| 총 | SST | $pqn-1$ | | |

분산 분석 - 이원배치 (반복O)

1. 가설 검정

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p$$

- 검정통계량 $F = \frac{MS_A}{MSE}$
- $F > F_\alpha(p-1, pq(n-1)) \rightarrow$ 귀무가설 기각

$$H_0: \gamma_{ij} = 0$$

- 검정통계량 $F = \frac{MS_{A,B}}{MSE}$
- $F > F_\alpha((p-1)(q-1), pq(n-1)) \rightarrow$ 귀무가설 기각

분산 분석 - 이원배치 (반복O)

1. 독립변수의 설명력

- A의 설명력: $\eta^2 = \frac{SS_A}{SST}$
- B의 설명력: $\eta^2 = \frac{SS_B}{SST}$
- 교호작용 설명력: $\eta^2 = \frac{SS_{A,B}}{SST}$
- A, B, AB 상호작용 설명력: $\eta^2 = \frac{SS_A + SS_B + SS_{A,B}}{SST}$
- 오차분산의 비율: $1 - \frac{SS_A + SS_B + SS_{A,B}}{SST}$

1. 회귀분석이란

- 독립변수와 종속변수 간 관련성을 설명하는 통계적 모형인 회귀모형을 통해, 두 변수의 데이터로 **회귀모형**에 적합한 **추정회귀식**을 계산하고 **통계적 분석(추론)**을 하는 기법
 - 회귀모형의 계수인 "**모수**"를 추정
 - 모수에 대한 구간추정/가설검정 등의 분석
 - 종속변수 설명에 있어서 독립변수의 상대적 중요성 평가 가능
- 예) 기업의 시가총액과 투자활동이 주식 수익률의 관련성 연구
- 과거의 데이터에 의존 (데이터 수집이 중요)
- 과거/현재/미래 예측

2. 회귀분석 종류

- 단순선형회귀분석
- 다중선형회귀분석
- 로지스틱회귀분석
- 비선형회귀분석

1. 회귀분석 종류

- 단순선형회귀모형

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

- 다중선형회귀모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- 비선형회귀모형

$$Y = \frac{m \exp(\beta_0 + \beta X)}{1 + \exp(\beta_0 + \beta X)} + \varepsilon$$

- 다변량회귀모형

$$\begin{cases} Y_1 = \beta_{10} + \beta_{11} X_1 + \beta_{12} X_2 + \cdots + \beta_{1p} X_p + \varepsilon \\ Y_2 = \beta_{20} + \beta_{21} X_1 + \beta_{22} X_2 + \cdots + \beta_{2p} X_p + \varepsilon \\ Y_3 = \beta_{30} + \beta_{31} X_1 + \beta_{32} X_2 + \cdots + \beta_{3p} X_p + \varepsilon \end{cases}$$

회귀 분석 - 단순회귀

1. 단순회귀모형

- 단순선형회귀모형

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- 추정 단순회귀식

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

2. 오차항 가정

- $E(\epsilon_i) = 0, \quad i = 1, \dots, n$
- 등분산: $\text{Var}(\epsilon_i) = \sigma^2, \quad i = 1, \dots, n$
- 독립성: $\text{Cov}(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$
- 정규성: $\epsilon_i \sim N(0, \sigma^2)$

회귀 분석 - 단순회귀

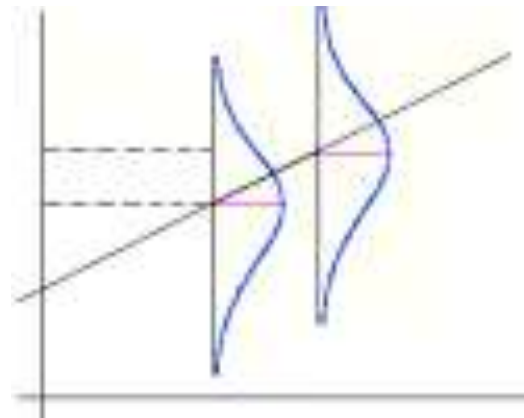
1. 종속변수의 분포

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\begin{aligned} E(y_i) &= E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i + E(\epsilon_i) \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

$$\begin{aligned} \text{Var}(y_i) &= \text{Var}(\beta_0 + \beta_1 x_i + \epsilon_i) = \text{Var}(\epsilon_i) \\ &= \sigma^2 \end{aligned}$$

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$



회귀 분석 - 단순회귀

1. 최소제곱법

- 어떤 관측값 y_i 에 대한 오차

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

- 모든 점에 대한 오차의 제곱합이 최소가 되도록 회귀계수를 추정

$$\sum \epsilon_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

2. 편미분 활용한 최소제곱법의 해

$$Q = \sum_i \epsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

회귀 분석 - 단순회귀

1. 편미분 활용한 최소제곱법의 해

$$Q = \sum_i \epsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (1)$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \quad (2)$$

(1), (2)에서 -2를 나눠주고 \sum_i 를 분배한 뒤, (1)에 $\bar{x} = \frac{\sum x_i}{n}$ 를 곱해준 뒤, (1)-(2)로 정리하면

$$\bar{x} \sum_i y_i - \sum_i x_i y_i + \beta_1 \sum_i x_i^2 - \bar{x} \beta_1 \sum_i x_i = 0$$

$$\beta_1 = \frac{\sum_i x_i y_i - \bar{x} \sum_i y_i}{\sum_i x_i^2 - \bar{x} \sum_i x_i} = \frac{\sum_i x_i y_i - 2\bar{x} \sum_i y_i + \bar{x} \sum_i y_i}{\sum_i x_i^2 - 2\bar{x} \sum_i x_i + \bar{x} \sum_i x_i} = \frac{\sum_i x_i y_i - 2\bar{x} \sum_i y_i + n\bar{x}\bar{y}}{\sum_i x_i^2 - 2\bar{x} \sum_i x_i + n\bar{x}^2} \quad (\sum_i x_i = n\bar{x} \text{를 활용})$$

$$= \frac{\sum_i x_i y_i - 2\bar{x} \sum_i y_i + \sum_i \bar{x}\bar{y}}{\sum_i x_i^2 - 2\bar{x} \sum_i x_i + \sum_i \bar{x}^2} = \frac{\sum_i (x_i y_i - 2\bar{x} y_i + \bar{x}\bar{y})}{\sum_i (x_i^2 - 2\bar{x} x_i + \bar{x}^2)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (\hat{\beta}_0 \text{는 } \hat{\beta}_1 \text{과 식(1)로부터 계산})$$

회귀 분석 - 단순회귀

1. 추정 회귀방정식

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

2. 적합값

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

3. 잔차

$$e_i = y_i - \hat{y}_i$$

- 잔차의 합은 0

회귀 분석 - 단순회귀

1. 잔차제곱합 (SSE)

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

2. MSE

$$MSE = \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

- 오차분산의 불편추정량
 - 오차분산 추정시 MSE 사용

$$E(MSE) = E\left(\frac{SSE}{n-2}\right) = \sigma^2$$

회귀 분석 - 단순회귀

1. β_1 의 추정 및 검정

$$\widehat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})y_i - \bar{y}\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

이 때, 분자의 $\sum (x_i - \bar{x})$ 는 편차의 합이므로 0이 됨 따라서,

$$= \frac{\sum (x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2} = \sum_i a_i y_i, \quad a_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$a_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$ 의 몇 가지 특징을 활용하여 회귀계수 추정값의 기댓값/분산 계산

$$1. \sum_i a_i = \sum_i \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) = \frac{\sum_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = 0$$

$$2. \sum_i a_i x_i = \frac{\sum_i (x_i - \bar{x})x_i}{\sum (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} = 1$$

회귀 분석 - 단순회귀

1. β_1 의 추정 및 검정

$$3. \sum_i a_i^2 = \sum \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 = \frac{1}{(\sum (x_i - \bar{x})^2)^2} \sum (x_i - \bar{x})^2 = \frac{1}{\sum (x_i - \bar{x})^2}$$

a_i 의 (1), (2) 특징을 활용하면 아래와 같이 기댓값을 구할 수 있음

$$E(\widehat{\beta_1}) = E(\sum a_i y_i) = \sum a_i E(y_i) = \sum a_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum a_i + \beta_1 \sum a_i x_i = \beta_1$$

a_i 의 (3) 특징을 활용하면 아래와 같이 분산을 구할 수 있음

$$\text{Var}(\widehat{\beta_1}) = \text{Var}(\sum a_i y_i) = \sum a_i^2 \text{Var}(y_i) = \sum a_i^2 \sigma^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

회귀 분석 - 단순회귀

1. β_1 의 추정 및 검정

$$V(\hat{\beta}_1) = V\left(\sum a_l y_l\right) = \sum a_l^2 V(y_l) = \sum a_l^2 \sigma^2 = \frac{1}{\sum (x_l - \bar{x})^2} \sigma^2 = \frac{\sigma^2}{S_{xx}}$$

$$V(\hat{\beta}_1) \text{의 불편추정량: } \frac{MSE}{S_{XX}} \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{XX}}} \sim t_{n-2}$$

β_1 의 100(1- α)% 신뢰구간

$$\left[\hat{\beta}_1 - t_{\left(\frac{\alpha}{2}, n-2\right)} \sqrt{\frac{MSE}{S_{XX}}}, \hat{\beta}_1 + t_{\left(\frac{\alpha}{2}, n-2\right)} \sqrt{MSE/S_{XX}} \right]$$

회귀 분석 - 단순회귀

1. β_1 의 추정 및 검정

- 검정통계량: $T = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{XX}}}$
- 양측검정
 - $H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$
- 단측검정
 - $H_0: \beta_1 = 0, H_1: \beta_1 > 0$
- 귀무가설 기각: β_1 이 0이 아니다.
즉, 독립변수 X는 종속변수에 대해 통계적으로 유의

1. β_0 의 추정 및 검정

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x}) = \frac{1}{n} \sum E(y_i) - \bar{x} E(\hat{\beta}_1)$$

$$= \frac{1}{n} \sum (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} = \beta_0$$

$$V(\hat{\beta}_0) = V(\bar{y} - \hat{\beta}_1 \bar{x}) = V(\bar{y}) + (\bar{x})^2 V(\hat{\beta}_1) - 2 \text{cov}(\bar{y}, \hat{\beta}_1 \bar{x})$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}} (\bar{x})^2 - 2\bar{x} \text{cov}(\bar{y}, \hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{1}{S_{xx}} (\bar{x})^2 \right) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

$$\begin{cases} \bar{y} = \sum a_i y_i & a_i = \frac{1}{n} \\ \hat{\beta}_1 = \sum c_i y_i & c_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{x_i - \bar{x}}{S_{xx}} \end{cases}$$

$$\Rightarrow \text{cov}(\bar{y}, \hat{\beta}_1) = \text{cov}(\sum a_i y_i, \sum c_i y_i) = \sigma^2 \sum a_i c_i = \frac{\sigma^2 \sum (x_i - \bar{x})}{n S_{xx}} = 0$$

(참고) y_i 가 독립일 때, $\text{cov}(\sum a_i y_i, \sum c_i y_i) = \sigma^2 \sum a_i c_i$

1. β_0 의 추정 및 검정

$$E(\hat{\beta}_0) = \beta_0$$

$$V(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

$$\Rightarrow \widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right)\right)$$

만약 오차분산 σ^2 를 모른다면

$$Var(\widehat{\beta}_0) = MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right)$$

회귀 분석 - 단순회귀

1. β_0 의 추정 및 검정

β_0 의 $100(1-\alpha)\%$ 신뢰구간

$$\left[\hat{\beta}_0 - t_{\left(\frac{\alpha}{2}, n-2\right)} \sqrt{MSE \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right]}, \hat{\beta}_0 + t_{\left(\frac{\alpha}{2}, n-2\right)} \sqrt{MSE \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right]} \right]$$

- 검정통계량: $T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right]}} \sim t(n-2)$

1. $E(Y_i)$ 의 추정 및 검정

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

$$E(\widehat{Y}_i) = E(\widehat{\beta}_0 + \widehat{\beta}_1 X_i) = \beta_0 + \beta_1 X_i = E(Y_i)$$

$$\begin{aligned} Var(\widehat{Y}_i) &= Var(\widehat{\beta}_0 + \widehat{\beta}_1 X_i) = Var\left(\bar{Y} + \widehat{\beta}_1(X_i - \bar{X})\right) \quad , (\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \text{ 활용}) \\ &= Var(\bar{Y}) + (X_i - \bar{X})^2 Var(\widehat{\beta}_1) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}} \right] \end{aligned}$$

회귀 분석 - 단순회귀

1. $E(Y_i)$ 의 추정 및 검정

- \hat{Y}_i 의 분포

$$\hat{Y}_i \sim iid N(E(Y_i), \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}} \right])$$

- 오차분산을 모를 때,

$$\hat{Y}_i \sim iid N(E(Y_i), MSE \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}} \right])$$

- $E(Y_i)$ 의 100(1- α)% 신뢰구간

$$\left[\hat{Y}_i - t_{\left(\frac{\alpha}{2}, n-2\right)} \sqrt{MSE \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}} \right]}, \hat{Y}_i + t_{\left(\frac{\alpha}{2}, n-2\right)} \sqrt{MSE \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}} \right]} \right]$$

1. $E(Y_i)$ 의 추정 및 검정

- 검정통계량: $T = \frac{\widehat{Y}_i - E(Y_i)}{S.E.(\widehat{Y}_i)} \sim t(n - 2)$
- 양측검정
 - $H_0: E(Y_i) = \mu_0, H_1: E(Y_i) \neq \mu_0$
- 귀무가설 기각: $E(Y_i)$ 이 μ_0 이 아니다.

회귀 분석 - 단순회귀

1. 변동 분해

- 총 변동 = 설명이 안되는 변동 + 설명이 되는 변동

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

- 총 제곱합(SST) = 잔차제곱합(SSE) + 회귀제곱합(SSR)

$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2$$

2. 분산분석

- 자유도: 총변동($n-1$) = 잔차변동($n-2$) + 회귀변동(1)
- 평균제곱

$$MSE = \frac{SSE}{n-2}, \quad MSR = \frac{SSR}{1}$$

회귀 분석 - 단순회귀

1. 회귀모형 검정 (F 검정)

- F 분포

$$F^* = \frac{MSR}{MSE} = \frac{SSR/\sigma^2}{\frac{SSR}{n-2}/\sigma^2} = \frac{\chi^2(1)/1}{\chi^2(n-2)/(n-2)} \sim F(1, n-2)$$

- F 분포와 t-분포

2. F-분포와 t-분포와의 관계

: $Z \sim N(0,1), V \sim \chi^2(k)$ 이고 Z 와 V 는 서로 독립일 때,

$$T = \frac{Z}{\sqrt{\frac{V}{k}}} \sim t(k), \quad T^2 = \frac{Z^2/1}{\frac{V}{k}} \sim F(1, k)$$

회귀 분석 - 단순회귀

1. 회귀모형 검정 (F 검정)

- 가설 검정

$$H_0: \beta_1 = 0, \quad H_1: \beta_1 \neq 0$$

- 검정통계량

$$F = \frac{MSR}{MSE}$$

- 검정

$$\begin{aligned} |F| > F_{(\alpha; 1, n-2)} & \text{ 귀무가설 기각} \\ |F| \leq F_{(\alpha; 1, n-2)} & \text{ 귀무가설 채택} \end{aligned}$$

1. 결정계수 (R^2)

- 총 변동 중 회귀식에 의해 설명되는 변동의 비율
 - 총 변동 중 어느 정도가 독립변수에 의해 설명되는가
- 0과 1사이의 범위에 있고 1에 가까울수록 회귀식에 의해 설명되는 양이 커짐

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \text{상관계수}^2$$