회귀분석입문 2주차 보충강의 : 기초통계학 review

* 통계적 추론

모집단
↓
표본추출
↓
모집단의 특성파악

모수 : 모집단의 특성

추정 : 모수를 표본을 통해 찾는 과정

검정 : 모수에 관한 가설을 표본을 통해 검증

* 확률표본 $\{X_1, X_2, X_3 \cdots X_n\}$: 가상의 이론 표본

모집단 $\sim N(\mu, \sigma^2) \longrightarrow X_i$는 서로 독립, $\sim N(\mu, \sigma^2)$

실제값 $\{x_1, x_2 \cdots x_n\}$

1. 모집단 $\sim N(\mu, \sigma^2)$ & σ^2이 알려진 경우 ($X_i \sim N(\mu, \sigma^2)$, 서로 독립)

1) μ의 점추정

① 추정량 : 표본평균 $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$

$E(\bar{X}) = \mu, \ Var(\bar{X}) = \dfrac{\sigma^2}{n}$ → μ에 대한 비편향추정량

② 추정값 : $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

실제 μ와 얼마나 가까운지 알수 없음

2) μ의 구간추정

$X_i \sim N(\mu, \sigma^2)$ 이고 서로 독립

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \longrightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ ; 표준정규분포}$$

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

↓

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

∴ μ의 95% 신뢰구간

$$\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}},\ \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)$$

⇒ 실제 μ를 포함하는 구간이 될 확률은 95%.

반복하여 구간을 구하면 1/20 꼴로 μ를 포함하지 않는 구간이 나온다는 뜻

$$\begin{pmatrix} 1.645 \to 0.05 \to 90\% \\ 1.96 \to 0.025 \to 95\% \\ 2.575 \to 0.005 \to 99\% \end{pmatrix}$$

3) μ의 가설검정

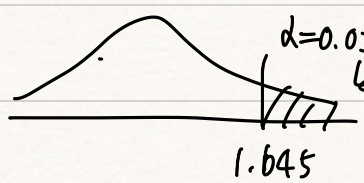
- H_0 귀무가설 : $\mu=\mu_0$
- H_1 대립가설
  - $\mu>\mu_0$, $\mu<\mu_0$ → 단측가설 : 단측검정
  - $\mu\neq\mu_0$ → 양측가설 : 양측검정

H_0 : $\mu=\mu_0$이 사실이라면 $Z=\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$은 사실

$P(Z\geq 1.645)=P(Z\leq -1.645)=0.05$

$P(Z\geq 1.96)=P(Z\leq -1.96)=0.025$ → $P(|Z|\geq 1.96)=0.05$

① $H_0:\mu=\mu_0$, $H_1:\mu>\mu_0$ if H_0 true → $P\left(\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}\geq 1.645\right)=0.05$



기각역 (critical region)

$$Z^*=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\geq 1.645$$

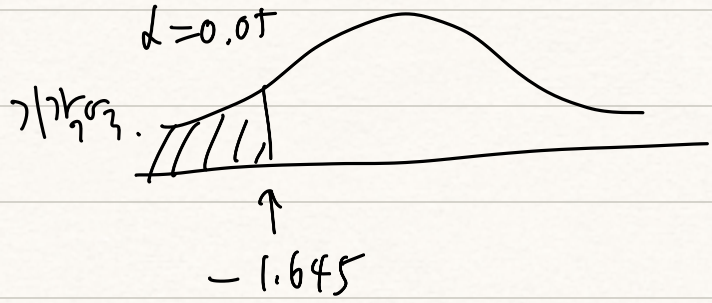
or $\bar{x}\geq \mu_0+1.645\dfrac{\sigma}{\sqrt{n}}$

귀무가설을 기각하는 확률

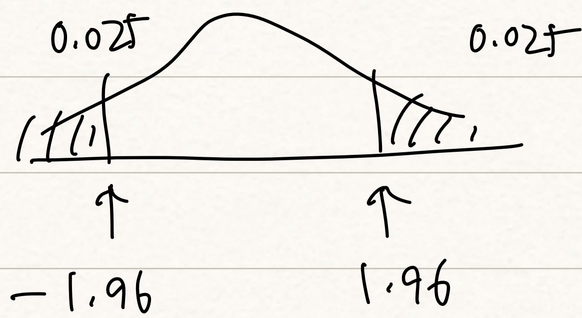
유의수준 $\alpha=0.05$의 검정 : H_0이 사실일때 $P(Z^*\geq 1.645)=0.05$

② $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$

$$z^* \le \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \le -1.645$$ 면 H_0 기각



③ $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \ne \mu_0 \end{cases}$



$$\Rightarrow \quad |z^*| = \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \ge 1.96$$ 인 경우 H_0 기각.

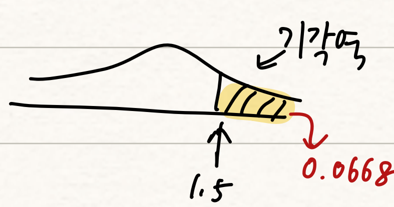
($\alpha = 0.05$).

ㅂ) 유의확률 (p-value)

H_0을 기각할 수 있게 되는 가장 작은 유의수준 α의 값

ex) $\begin{cases} H_0 : \mu = 100 \\ H_1 : \mu > 100 \end{cases}$ $\bar{x} = 101.5$, $\sigma^2 = 25$. $n = 25$ 인 경우

$$z^* = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{101.5 - 100}{5/5} = 1.5$$



$$p\text{-value} = P(Z \ge 1.5) = 0.0668$$

$$\left(\begin{array}{l} \text{단측} \rightarrow \text{p-value} = P(Z \geq z^*) \\ \text{양측} \rightarrow \text{p-value} = 2 \cdot P(Z \geq z^*) \end{array}\right).$$

2. 모집단 $\sim N(\mu, \sigma^2)$ & σ^2이 알려지지 않은경우 ($X_i \sim N(\mu, \sigma^2)$, 서로독립)

1) μ의 점추정

μ 추정량 : $\bar{X}$

μ 추정값 : $\bar{x}$

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

σ^2 추정값 $s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$

2) μ의 구간추정.

$X_i \sim N(\mu, \sigma^2)$

σ^2를 S^2로 대체 $\rightarrow \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

자유도가 n-1인 T분포

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

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

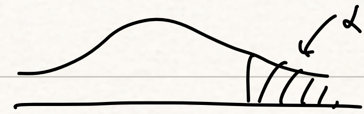
3) μ의 가설검정

$H_0 : \mu = \mu_0 \rightarrow \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$

$$\left(\begin{array}{l} \sigma^2 \longrightarrow s^2 \\ N(0,1) \longrightarrow t(n-1) \end{array}\right)$$

①
$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$$

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq t(\alpha, n-1)$$ 이면 기각



$$p\text{-value} = P(t(n-1) \geq t^*)$$

②
$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$$

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq t(\alpha, n-1)$$ 이면 기각

$$p\text{-value} = P(t(n-1) \leq t^*)$$

③
$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$$

$$|t^*| = \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| \geq t(\alpha/2, n-1)$$ 이면 기각

$$p\text{-value} = 2P(t(n-1) \geq |t^*|)$$

< 분포이론 >

- 정규분포

$X \sim N(\mu, \sigma^2)$

$E(X) = \mu,\ Var(X) = \sigma^2$

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < \infty$$

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$aX + b \sim N(a\mu + b,\ a^2\sigma^2)$$

- χ^2-분포

z_i가 표준정규분포를 따르고 서로 독립인 경우

$$\begin{cases} z_1^2 \sim \chi^2(1) \\ z_1^2 + z_2^2 \sim \chi^2(2) \\ \quad \vdots \\ \sum_{i=1}^{k} z_i^2 \sim \chi^2(k) \end{cases}$$

- t-분포

$z \sim N(0,1),\ u \sim \chi^2(k)$, z와 u는 서로 독립

$$\frac{z}{\sqrt{u/k}} \sim t(k)$$

· F-분포

$u \sim \chi^2(k_1)$, $v \sim \chi^2(k_2)$, u와 v는 서로 독립

$$\frac{u/k_1}{v/k_2} \sim F(k_1, k_2)$$

$u \sim t(k)$

$u^2 \sim F(1,k)$

ex) $t(0.025, 6) = 2.447$

자유도가 6인 t-분포에서 2.447보다 클 확률은 0.025

$\rightarrow \left(t(0.025, 6)\right)^2 = 5.99$, $F(0.05, 1, 6) = 5.99$

제곱이므로 t분포가 2.447보다 큰경우/작은경우 모두 고려

<확률변수>

· 기댓값

이산형 Y : $E(Y) = \sum y \cdot f(y) = \mu$

연속형 Y : $E(Y) = \int y \cdot f(y)\, dy = \mu$

· 분산

$$\sigma^2 = Var(Y) = E(Y - E(Y))^2 = E(Y-\mu)^2 = E(Y^2) - \mu^2$$

· 공분산

$$\sigma_{12} = Cov(Y_1 \cdot Y_2) = E(Y_1-\mu_1)(Y_2-\mu_2) = E(Y_1Y_2) - \mu_1\mu_2$$

· 상관계수

$$\rho_{12} = Corr(Y_1 . Y_2) = \frac{Cov(Y_1 . Y_2)}{\sqrt{Var(Y_1)}\sqrt{Var(Y_2)}}$$

$E(a) = a$ (상수)

$E(aY+b) = aE(Y)+b$

$Var(aY+b) = a^2 Var(Y)$

$E(a_1Y_2 + a_2Y_2) = a_1E(Y_1) + a_2E(Y_2)$

$Var(a_1Y_1 + a_2Y_2) = a_1^2 Var(Y_1) + a_2^2 Var(Y_2) + 2a_1a_2Cov(Y_1 . Y_2)$

$Cov(a_1Y_1 + b_1, a_2Y_2 + b_2) = a_1a_2 Cov(Y_1 . Y_2)$

$Corr(a_1Y_1 + b_1, a_2Y_2 + b_2) = sign\{(a_1a_2)\}\ Corr(Y_1 . Y_2)$

↳ a_1a_2가 양수 → 그대로

a_1a_2가 음수 → 부호바꾸기

n개의 확률변수 $Y_1, Y_2, \cdots Y_n$

$a_1Y_1 + a_2Y_2 + \cdots + a_nY_n = \sum_{i=1}^{n} a_iY_i$; linear combination

- $E\left(\sum_{i=1}^{n} a_iY_i\right) = E(a_1Y_1 + \cdots + a_nY_n)$

$$= a_1E(Y_1) + a_2E(Y_2) + \cdots + a_nE(Y_n)$$

$$= \sum_{i=1}^{n} a_iE(Y_i)$$

- $Var\left(\sum_{i=1}^{n} a_iY_i\right) = \sum_{i=1}^{n} a_i^2 Var(Y_i) + \sum_{i \neq j} a_ia_j Cov(Y_i, Y_j)$

$$= a_1^2 Var(Y_1) + a_2^2 Var(Y_2) + \cdots + a_n^2 Var(Y_n)$$

$$+ a_1a_2 Cov(Y_1, Y_2) + a_1a_3 Cov(Y_1, Y_3) + \cdots + a_{n-1}a_n Cov(Y_{n-1}, Y_n)$$

또는

$Var\left(\sum_{i=1}^{n} a_iY_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j Cov(Y_i, Y_j)$

$$= a_1a_1 Cov(Y_1, Y_1) + a_1a_2 Cov(Y_1, Y_2) + \cdots + a_na_n Cov(Y_n, Y_n)$$

$Cov(Y_i, Y_j) = 0$ 이라면

$$Var\left(\sum_{i=1}^{n} a_iY_i\right) = \sum_{i=1}^{n} a_i^2 Var(Y_i) + \underbrace{\sum_{i \neq j} a_ia_j Cov(Y_i, Y_j)}_{\to 0}$$

$$= \sum_{i=1}^{n} a_i^2 Var(Y_i)$$

$$= a_1^2 Var(Y_1) + a_2^2 Var(Y_2) + \cdots + a_n^2 Var(Y_n)$$

- $Cov\left(\sum_{i=1}^{n} a_iY_i, \sum_{j=1}^{m} b_jX_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_ib_j Cov(Y_i, X_j)$

- $Cov(Y_1 + Y_2, Y_3) = Cov(Y_1, Y_3) + Cov(Y_2, Y_3)$