

〈 분포 〉

- 순열 (permutation) : $n P_n = n(n-1)!$, r 개 순서 고려 : $n P_r = \frac{n!}{(n-r)!}$
- 조합 (combination) : $n C_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} = n C_{n-r}$
- 조건부 확률 : B 사건 가정 하 A 사건의 확률 $= P_B(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}$
- 베이즈 공식 : $P(A_i | B) = \frac{P(A_i) P(B|A_i)}{P(A_1) P(B|A_1) + \dots + P(A_n) P(B|A_n)}$
- ex) $\begin{matrix} B & \\ \text{불량품이} & A_1 \\ \text{1 호기에서} & \end{matrix}$ 생산되었을 확률 \Rightarrow 불량품을 뽑았는데 ① \rightarrow 그게 1호다 ②
 $\Rightarrow P(A_1 | B)$
- 베르누이 시행 : $f(x) = p^x (1-p)^{1-x}$, $x = 0, 1$, $0 \leq p \leq 1$... $X \sim B(1, p)$
- 이항분포 : $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \dots, n$, ... $X \sim B(n, p)$
- 포아송 분포 : $f(x) = \frac{e^{-\lambda} \lambda^x}{x!}$, $x = 0, 1, 2, \dots$, λ : 단위 내 발생하는 사건의 평균값
 $e = \lim_{n \rightarrow \infty} (1 + \frac{\lambda}{n})^n = 2.71828 \dots$, $X \sim P(\lambda)$
- 기하분포 : 성공 확률 p 의 베르누이 분포. 차음으로 성공할 때 까지의 시행 횟수 확률 변수 X ,
 $f(x) = pq^{x-1}$, $x = 1, 2, 3, \dots$, $X \sim G(p)$
- 음이항분포 : 성공 확률 p 의 베르누이 분포, k 번 성공할 때 까지의 시행 횟수 확률 변수 X ,
 $f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$, $x = k, k+1, k+2, \dots$

$X \sim NB(k, p)$, k 번 성공할 때 까지의 실패 횟수 확률변수 X ,

$$f(x) = \binom{x+k-1}{x} p^k (1-p)^x, \quad x = 0, 1, 2, \dots$$

◦ 초기하분포 : 유한모집단 N 중 크기 n 의 확률표본을 뽑을 경우, N 개 중 k 개는 성공으로,

나머지 $(N-k)$ 개는 실패로 분류하여 비복원 추출로 뽑을 때의 성공 횟수 X ,

$$f(x) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2, \dots, n, \quad X \sim HG(N, k, n)$$

◦ 지수분포 : 한 번의 사건이 발생할 때 까지의 소요되는 시간의 분포, 한번 사건이 발생할 때 까지의 소요시간 X , 단위시간 당 평균 사건 횟수 λ ,

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \quad \lambda > 0 \quad \dots \quad X \sim E(\lambda)$$

◦ 감마분포 : α 번의 사건이 발생할 때 까지의 대기시간의 분포

$$f(x) = \frac{1}{T(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} \quad \dots \quad X \sim T(\alpha, \beta)$$

◦ 카이제곱분포는 보분산 σ^2 이 특정한 값을 갖는지 여부를 검정하는데 사용되는 분포이며, 두 범주형 변수 간의 연관성을 검정하는데 주로 사용된다. 카이제곱분포의 기대값은 자유도이며, 분산은 자유도의 두 배이다.

- 자유도란 독립적인 관측값의 개수를 의미한다.

- t 분포는 자유도가 증가함에 따라 표준정규분포에 수렴한다. t 분포는 소표본에 주로 이용되는 분포이다.
- F 분포는 분산분석과 회귀분석에서 집단간 분산비 검정에 주로 사용한다.
- $T \sim t(n)$ 일 때, $T^2 \sim F(1, n)$ 이다.

- X 가 $F(m, n)$ 을 따를 때 $\frac{1}{X}$ 의 분포는 $F(n, m)$ 을 따른다.
- F 분포는 두 집단의 분산비 검정, 세 집단 이상의 모평균 비교에 사용된다.
- t 분포, F 분포, χ^2 분포는 자유도에 의존해 확률을 구한다.
- 이항분포의 정규근사 : $B(n, p)$ 에서 $np > 5$ 이고 $n(1-p) > 5$ 이면 $B(n, p)$ 는 정규분포 $N(np, npq)$ 에 근사한다. $P(a < X < b) \approx P\left(\frac{a-np}{\sqrt{npq}} < \frac{X-np}{\sqrt{npq}} < \frac{b-np}{\sqrt{npq}}\right)$
하지만 n 이 크더라도 p 가 0에 가까운 경우 이항분포보다 포아송분포에 근사시키는 것이 더 정확하다.

〈 추정 〉

- 유한모집단 (비복원추출) 시 표본평균의 표본분포 : 표본평균 \bar{X} 의 기대값 : $E(\bar{X}) = \mu$
- * 분산 $\sigma_{\bar{X}}^2 = \frac{N-n}{N-1} \frac{\sigma^2}{n}$, $\frac{N-n}{N-1}$: 유한모집단 수정계수 분산 : $Var(\bar{X}) = \sigma_{\bar{X}}^2$
- 무한모집단 (복원추출) 시 표본평균의 표본분포 : \bar{X} 의 기대값 : $E(\bar{X}) = \mu$
분산 : $Var(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$
표준오차 : $\sqrt{Var(\bar{X})} = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- 표준편차 : $\sigma = \sqrt{\frac{1}{N} \sum_1^n (x_i - \mu)^2}$ • 표본표준편차 : $s = \sqrt{\frac{1}{n-1} \sum_1^n (x - \bar{x})^2}$
- 표준오차 : 추정량의 표준편차 . 표본평균의 표준오차 : $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
표본비율의 표준오차 : $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

〈통계량〉

- 사분위 범위 $IQR = Q_3 - Q_1$, 얀울타리 $IL = Q_1 - (1.5 \times IQR)$

$$IU = Q_3 + (1.5 \times IQR), \text{ 바깥울타리 } OL = Q_1 - (3 \times IQR), OU = Q_3 + (3 \times IQR)$$

- 기하평균 (시간적으로 변화하는 비율) : $GM = \sqrt[n]{\prod x_i}$, $\sqrt[n]{x_1 \times x_2 \times \dots}$

- 조화평균 (시간적으로 변화하는 속도) : $HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots}$, $n=2, \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$

- 평방평균 : $QM = \sqrt{\frac{1}{n} \sum x_i^2}$

- 절사평균 : 앞 두 $n\%$ 씩 없애고 산술평균 냈것.

- 사분 편차 (Quanfile Deviation) = $\frac{Q_3 - Q_1}{2}$

- 사분위 편차계수 (Coef of Quantile Dev) = $\frac{QD}{m}$, m = 중앙값

- 변동계수 (Coef Variance) : ① 모집단 = $\frac{\sigma}{\mu}$ ② 표본집단 = $\frac{s}{\bar{x}}$

- 상대분산 = CV^2 : 평균에 대한 표준편차의 상대적 크기를 비교할 때 사용

- 표준화 점수 : ① 모집단 = $Z = \frac{X - \mu}{\sigma}$ ② 표본집단 = $Z = \frac{X - \bar{X}}{S_x}$

- 공분산 : 두 변수 간의 선형연관성을 나타내는 측도. $Cov(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$

- 상관계수 : 공분산을 두 확률변수의 표준편차의 곱으로 나눈 값. $r = \frac{Cov(X, Y)}{S_x S_y} \Rightarrow$ 모수적 통계분석

- 피어슨 상관계수 (표본상관계수), 스피어만 상관계수 (순위상관계수) \rightarrow 비모수적 통계분석 + (肯德尔의 타우 검정)
(비모수) ~

- Kruskal-Wallis test, Mann-Whitney test, Wilcoxon Signed Rank test, Kendall's Tau test

- 적률 : 함수의 모양을 수학적으로 표현하는 하나의 척도

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx \quad (\text{보통 } c=0)$$

- 평균 (1차 적률), 분산 (2차 적률), 왜도 (3차 적률), 첨도 (4차 적률)

$$\mu'_n = \int_{-\infty}^{\infty} x^n f(x) dx \quad \dots \text{ } n\text{ 차 적률은 } X^n \text{ 의 평균. } E[X^n]$$

• 모비율 검정통계량 $Z = \frac{\hat{P} - P_0}{\sqrt{\hat{P}(1-\hat{P})/n}}$ 모비율이 알려져 있으면
분모의 \hat{P} 가 모비율로 대체됨

• 모비율이 알려져 있으면 $Z = \frac{\hat{P} - P_0}{\sqrt{P_0(1-P_0)/n}}$

- t 검정의 기본 가정 : ① 종속 변수가 양적 변수 ② 모집단의 분산을 알지 못함
③ 모집단의 분포가 정규분포 ④ 등분산 가정 충족

- 두 모분산을 알고 있을 경우 두 모평균의 차 $\mu_x - \mu_y$ 에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\left((\bar{X} - \bar{Y}) - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}, (\bar{X} - \bar{Y}) + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right)$$

- * 신뢰구간에 0이 포함되어 있으면 두 평균 간에 차이가 없다는 귀무가설을 채택한다.
- * 새롭게 주장하는 가설 : H₁, 대립가설, 기존에 믿어지고 있는 가설 : H₀, 귀무가설

- 모집단이 정규분포인 대표본에서 모분산의 검정은 χ^2 검정을 이용한다.
- * 독립시행의 수가 n , 성공비율이 θ 인 이항모집단에서 성공 수 X 를 관측시,

$H_0 : \theta = p$, $H_1 : \theta = 1-p$ 라 할 때, P-value 는?

$$= \sum_{i=x}^n \binom{n}{i} p^i (1-p)^{n-i}$$
- 중심극한정리 : 표본의 크기가 $n \geq 30$ 이면 대표본으로 간주하여 모집단의 분포와 관계없이 표본평균 \bar{X} 의 분포는 기대값이 모평균 μ 이고, 분산이 $\frac{\sigma^2}{n}$ 인 정규분포에 근사한다.
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
- 체비체프 부등식 : 확률변수 X 에 대해 $E(X) = \mu$ 이고 $Var(X) = \sigma^2$ 일 때, 임의의 양수 k 에 대해 다음이 성립한다.
$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \quad \text{또는} \quad P(|X - \mu| \leq k) \geq 1 - \frac{\sigma^2}{k^2}$$
- 정규분포 : 평균 μ , 분산 σ^2 ... 정규분포 $N(\mu, \sigma^2)$ 를 따를 때
$$P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Φ : 누적표준 정규분포

- 표본평균의 분산 : $X \sim N(\mu, \sigma^2)$ 일 때, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ 을 따르므로
 (\bar{X}) $Var(\bar{X}) = \frac{\sigma^2}{n}$, 표준오차 = $\frac{\sigma}{\sqrt{n}}$
- 모분산 σ^2 의 추정량은 표본분산 S^2 이며 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$ 을 따른다는 χ^2 분포를 이용.

- 대표본에서 두 모분산을 알고 있을 때 $\mu_1 - \mu_2$ 에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\therefore (\bar{X}_A - \bar{X}_B) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

$$\circ \text{ 표본비율의 표준오차 } \hat{\sigma} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\circ \text{ 표본비율의 평균제곱오차 } \text{MSE}(\hat{p}) = \text{Var}(\hat{p}) + [\text{Bias}(\hat{p})]^2$$

$$\circ \text{ 모분산비 } \frac{\sigma_2^2}{\sigma_1^2} \text{에 대한 } 100(1-\alpha)\% \text{ 신뢰구간은 다음과 같다.}$$

$$\left(F_{1-\frac{\alpha}{2}, m-1, n-1} \frac{S_2^2}{S_1^2}, F_{\frac{\alpha}{2}, m-1, n-1} \frac{S_2^2}{S_1^2} \right)$$

- 소표본에서 두 모분산을 모르지만 같다는 것을 알고 있는 경우 $100(1-\alpha)\%$ 신뢰구간

t 통계량 사용 ...

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

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

$$\therefore \left((\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}, (n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

$$\circ \text{ 분산비 } \frac{\sigma_1^2}{\sigma_2^2} \text{에 대한 신뢰구간} : \left(\frac{1}{F_{\frac{\alpha}{2}, \phi_1, \phi_2}} \cdot \frac{S_1^2}{S_2^2}, F_{\frac{\alpha}{2}, \phi_2, \phi_1} \cdot \frac{S_1^2}{S_2^2} \right)$$

$$\phi_1 = m-1, \phi_2 = n-1$$

◦ 최대가능도 추정량 (MLE) : $L(\theta) = L(\theta; x) = f(x_1, \dots, x_n; \theta)$

$$= f(x_1; \theta) \cdot f(x_2; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

다공해서 로그 씀다음에
θ에 대해 미분.

◦ 적률법 : 적률추정량 MME (method of moments estimator)

모집단의 r차 적률을 $\mu_r = E(X^r)$ 이라 하고 표본의 r차 적률을 $\hat{\mu}_r = \frac{1}{n} \sum x_i^r$ 이라 할 때 모집단의 적률과 표본의 적률을 같다 ($\mu_r = \hat{\mu}_r$) 고 놓고 모수에 대해 추정량을 구하는 방법.

* 데이터 분포의 첨도가 10 이상이면 정규성 가정을 만족한다고 할 수 없다.

◦ 바람직한 추정량의 성질

① 불편성 (Unbiasedness) : 모수 θ 의 추정량을 $\hat{\theta}$ 으로 나타낼 때, $\hat{\theta}$ 의 기대값이 θ 가 되는 성질. 즉, $E(\hat{\theta}) = \theta$ 이면 $\hat{\theta}$ 을 불편추정량이라 한다.

② 일치성 (Consistency) : 표본의 크기가 커짐에 따라 추정량 $\hat{\theta}$ 이 확률적으로 모수 θ 에 가깝게 수렴하는 성질이다.

③ 충분성 (Sufficiency) : 모수에 대하여 가능한 많은 표본 정보를 대포하고 있는 추정량의 성질.

④ 효율성 (Efficiency) : 추정량 $\hat{\theta}$ 이 불편추정량이고, 그 분산이 다른 추정량 $\hat{\theta}_i$ 에 비해 최소의 분산을 갖는 성질이다.

◦ 점추정 방법 : ① 적률법 (Method of Moment) MME

② 최대가능도 추정법 (Method of Maximum Likelihood) MLE

Maximize $L(\theta) = f(x_1, \dots, x_n; \theta)$ $\hat{\theta}$... log 써어서 로그 가능도 함수

$\log L(\theta)$ = MLE로 주로 사용.

◦ 평균제곱오차 (MSE) = $E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$

→ 추정량의 분산 + 편차제곱

... 비편향이면서 분산이 큰 추정량 VS 편향이면서 분산이 작은 추정량 문제의 판단 기준이 됨.

◦ 구간추정 ... 신뢰구간 : 일정한 구간을 제시하여 모수가 포함되었을 것이라고 제시한 구간

신뢰수준 : 신뢰수준 95% 라 함은 같은 연구를 100 번 반복해서 신뢰구간을
구하는 경우, 그 중 적어도 95 번은 그 구간안에 모평균이 포함됨.

◦ 신뢰계수 : 추정량의 분포와 신뢰수준에 의해 결정

- μ 의 90% 신뢰구간 = $\bar{X} \pm 1.645 \frac{\sigma}{\sqrt{n}}$

- μ 의 95% 신뢰구간 = $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$

- μ 의 99% 신뢰구간 = $\bar{X} \pm 2.575 \frac{\sigma}{\sqrt{n}}$

◦ 90, 95, 99% 를 신뢰수준 이라고 하며, $Z_{0.05} = 1.645$, $Z_{0.025} = 1.96$,

$Z_{0.005} = 2.575$ 를 신뢰계수 라 하며, $\frac{\sigma}{\sqrt{n}}$ 를 표준오차 라 한다.

◦ 같은 신뢰수준에서 신뢰구간이 좁다는 것은 추정이 정밀하게 되었다는 뜻이다.

$100(1-\alpha)\%$

◦ 표본 크기 결정 : 모평균의 추정에 필요한 표본 크기 결정

.. X_1, X_2, \dots, X_n 은 평균이 μ , 분산이 σ^2 인 오차단에서의 확률표본일 때 모평균 μ 의

$100(1-\alpha)\%$ 신뢰구간은 $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ 이다. $\frac{\sigma}{\sqrt{n}}$ 를 표준오차라 하고,

$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ 을 추정오차(오차한계)라 하며, 추정 오차가 d 이내가 되도록 하려면

$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq d$ 로부터 다음과 같이 표본크기 n 을 결정 가능하다.

$$n \geq \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{d} \right)^2$$

◦ 모비율의 추정에 필요한 표본 크기 결정

모비율 p 에 대한 $100(1-\alpha)\%$ 신뢰구간은 $\bar{X} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ 이다.

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$: 표준오차, $z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$: 추정오차(오차한계)라 할 때 $\leq d$

$$\Rightarrow n \geq \hat{p}(1-\hat{p}) \left(\frac{z_{\frac{\alpha}{2}}}{d} \right)^2$$

◦ 모비율 p 에 대한 과거의 경험이나 시설조사를 통해 사전정보가 있을 경우,

표본의 크기는 $n \geq \hat{p}(1-\hat{p}) \left(\frac{z_{\frac{\alpha}{2}}}{d} \right)^2$ 을 사용하지만, 모비율 p 에 대한

사전정보가 없을 경우에는 보수적인 방법으로 $\hat{p}(1-\hat{p})$ 을 최대로 하는 $\hat{p} = \frac{1}{2}$ 을 대입하여 표본의 크기를 결정한다.

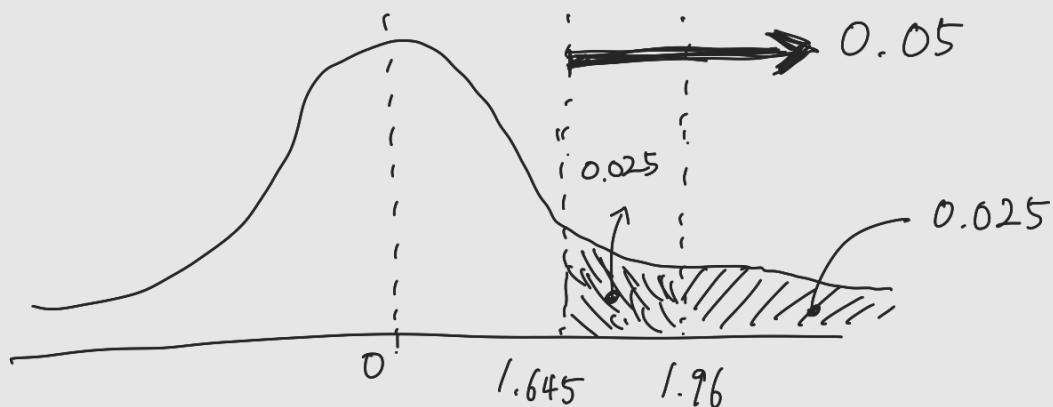
• 표본 분산을 구할 때 왜 $n-1$ 로 나누는가?

- ① 표본의 분산은 모집단의 분산을 과소 평가하여 표본 분산 < 모집단 분산 같은 상태가 되기 때문에 이를 보정해주기 위해 표본 분산의 분모를 작게 한다.
- ② 샘플 분산에서는 자유도가 $n-1$ 이기 때문이다.
- ③ 표본 분산의 기대값을 구할 경우 수학적으로 정확히 모집단의 분산으로 유도되기 때문이다.
 $n-1$ 로 나누어야 딱 맞아떨어진다.

- 모집단이 정규분포인 대표본에서 모분산의 검정은 χ^2 검정을 이용한다.
- 독립시행의 수가 n , 성공 비율이 θ 인 이항모집단에서 성공 수 X 를 관측시,
 $H_0: \theta = P$, $H_1: \theta = 1-P$ 라 할 때, $p\text{-value}$ 는 다음과 같다.

$$p\text{-value} = \sum_{i=x}^n \binom{n}{i} p^i (1-p)^{n-i}$$

- 결정 원칙 : 검정통계량값 (Z_0), 기각치 (Z_α) $P(Z > 1.96) = 0.025$
- 좌측 검정 : $Z_0 \leq -Z_\alpha$ 이면 귀무가설 기각 $P(Z > 1.645) = 0.05$
- 우측 검정 : $Z_0 \geq Z_\alpha$ 이면 귀무가설 기각
- 양측 검정 : $|Z_0| \geq Z_{\frac{\alpha}{2}}$ 이면 귀무가설 기각



• 모비율 σ^2 을 모르고 있을 경우 모평균 μ 에 대한 검정통계량은

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \text{ 을 따른다.}$$

\hookrightarrow 자유도

• 두 모비율 차 $p_1 - p_2$ 에 대한 검정 ... $H_0 : p_1 = p_2$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{합동표본비율 } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

• 소표본에서 두 모분산을 모르지만 같라는 것을 놓는 경우 : 두 모평균의 차 $\mu_1 - \mu_2$ 검정

$$H_0 : \mu_1 = \mu_2, \quad t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{합동분산 } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$

• 대응표본 t 검정의 검정통계량 : $t = \frac{\bar{D}}{S_D / \sqrt{n}}$

• 대표본에서 두 모분산을 모르고 있을 경우 검정통계량 $Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

• 모분산 검정에 사용되는 검정통계량은 $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$ 이다.
(카이제곱검정)

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

< 가설 검정 >

- 단측검정 : 사건 정보에 의해 표본분포의 한쪽에 관심을 가지고 시행하는 검정 방법
- 양측검정 : 사건 정보가 없거나 표본분포의 양쪽에 관심을 가지고 시행하는 검정 방법
- 1종 오류 : 귀무가설이 옳은데도 대립가설을 채택 : α 1종 오류를 범할 확률 : 유의수준
 ↓
• 2종 오류 : 대립가설이 옳은데도 귀무가설을 채택 : β 귀무가설이 참일 때, 대립가설을 채택하는 오류를 범할 확률
- 1종 오류와 2종 오류는 상호 역의 관계에 있으므로 하나를 줄이면 다른 하나가 증가한다.
- 독립표본 t 검정 : ① 조사 대상 개체가 다름 ② 두 표본의 숫자가 다를 수 있음
(다른 집단) ③ 다른 집단을 비교하는 경우 ④ 두 표본이 서로 독립
- 대응표본 t 검정 : ① 조사 대상 개체가 같음 ② 반드시 짝을 이룸
(같은 집단 내) ③ 전·후 개념이 있는 경우가 많음 ④ 두 표본이 서로 독립이 아님
- * 결정 원칙 : 가설의 기각 여부는 귀무가설을 기준으로 생각하지만 결론은 대립가설을 기준으로 서술.
- 좌측검정인 경우 ($H_0 : \mu_1 = \mu_0$, $H_1 : \mu_1 < \mu_0$) , $P(\bar{X} > \bar{x}_{obs})$
- 우측검정인 경우 ($H_0 : \mu_1 = \mu_0$, $H_1 : \mu_1 > \mu_0$) , $P(\bar{X} < \bar{x}_{obs})$
- 유의확률 ($p\text{-value}$) 과 유의수준 (α) 을 비교하여 $p\text{-value} < \alpha$ 이면 귀무가설을 기각한다. (단측)
- 양측 검정인 경우 ($H_0 : \mu_1 = \mu_0$, $H_1 : \mu_1 \neq \mu_0$) , $P(|\bar{X}| > |\bar{x}_{obs}|)$

* 양측 검정의 경우 P-value 는 단측 검정 p-value 의 2배가 된다.

• 모분산 σ^2 을 알고 있는 경우 모평균 μ 의 검정

가설: $H_0 : \mu = \mu_0$, 검정통계량 $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

이 외의 다양한 경우는
책 302p 참조

• 모분산 σ^2 을 모르고 있을 경우 모평균 μ 의 검정

가설: $H_0 : \mu = \mu_0$, 검정통계량 $t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$

• 모분산 $\sigma_1^2 = \sigma_2^2$ 에 대한 검정을 Levene 의 등분산 검정이라 하며, 분석 결과

F 값의 p-value 가 유의수준보다 작아 등분산이 가정되지 않으면 t 검정 결과를 이용해야 한다.

• 검정력 (power) : 전제 확률 |에서 2종 오류를 범할 확률 β 를 뺀 확률로,
 $<1-\beta>$ 대립가설이 참일 때 귀무가설을 기각할 확률이다.

제 1종 오류 α 를 고정시킨 상태에서 표본의 크기를 증가시키면
 β 를 감소시킬 수 있다 \rightarrow 검정력이 증가한다.

같은 유의수준과 같은 표본 크기에서는 검정력이 큰 검정법이 더 좋다.

• 유의확률 (p-value) : p 값이란 귀무가설이 사실이라는 전제 하에 검정통계량이 표본에서 계산된 값과 같거나 그 값보다 대립가설 방향으로 더 극단적인 값을 가질 확률이다.

즉, p 값은 검정통계량 값에 대해서 귀무가설을 기각시킬 수 있는 최소의 유의수준으로 귀무가설이 사실일 확률이라 생각할 수 있다.

- 유의수준 (α) : 유의 수준은 귀무가설이 참일 때 귀무가설을 기각할 오류 (제 1종 오류)를 범할 확률의 최대허용한계를 의미한다. 즉, 유의수준 0.05는 제 1종 오류를 범할 확률을 최대 5% 까지는 허용하여, 5%를 초과할 경우 허용하지 않겠다는 의미이다.

- 비율에 대한 검정통계량 $Z = \frac{\hat{P} - P_0}{\sqrt{\hat{P}(1-\hat{P})/n}}$... 모비율이 알려져있으면 본모의 \hat{P} 가 모비율로 대체됨.

$$\rightarrow \text{모비율 } P \text{ 를 알때, } Z = \frac{\hat{P} - P_0}{\sqrt{P_0(1-P_0)/n}}$$

- t 검정의 기본 가정 :
 - ① 종속 변수가 양적 변수
 - ② 모집단의 분산을 모를 때 사용
 - ③ 모집단의 분포가 정규분포
 - ④ 등분산 가정 충족

- 두 모분산을 알고 있을 경우 두 모평균의 차 $\mu_X - \mu_Y$ 에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\left((\bar{X} - \bar{Y}) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}, (\bar{X} - \bar{Y}) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right)$$

- 신뢰구간에 0이 포함되어 있으면 두 ^(평균) ~ 간에 차이가 없다는 귀무가설을 채택한다.

- 새롭게 주장하는 가설 : 대립가설 (H_1), 기존에 믿고 있는 가설 : 귀무가설 (H_0)

〈 범주형 자료분석 〉

- 두 범주형 변수 간의 연관성, 관련성을 검정하는 경우 카이제곱 χ^2 검정을 이용한다.
교차분석이라고도 하며 카이제곱 독립성 검정과 카이제곱 동일성 검정으로 나뉜다.
비모수적 방법으로 카이제곱 적합성 검정이 있다.
- 오즈비 (odds ratio) : 오즈는 사건이 발생하지 않은 확률에 대한 사건이 발생한 확률의 비율이다. 오즈비는 각각의 오즈에 대한 비율을 의미한다.
$$\text{odds} = \frac{\text{사건 발생 확률}}{\text{사건 미발생 확률}}$$
 계산된 오즈비가 3이라면 A 그룹에서 사건이 발생할 확률은 B 그룹에서 사건이 발생할 확률의 3배라고 해석할 수 있다.
- 카이제곱 독립성 검정 : 두 범주형 변수 간에 서로 연관성이 있는지를 검정
ex) 성별과 흡연 사이의 연관
 H_0 : A 와 B 는 서로 독립이다. H_1 : A 와 B 는 서로 독립이 아니다.
유의수준 α 일 때 결과 해석 : $\chi^2 \geq \chi^2_{\alpha, (r-1)(c-1)}$ \Rightarrow 귀무가설 기각 : 독립이 아니다
 $\chi^2 < \chi^2_{\alpha, (r-1)(c-1)}$ \Rightarrow 귀무가설 채택 : 독립이다
- 카이제곱 동일성 검정 : 하나의 특성에 대하여 몇 개의 범주로 분류된 자료가 주어졌을 때
ex) 지역에 따른 흡연 정도 차이 여러 모집단들이 주어진 특성에 대하여 서로 동일한 분포를 하는지 검정
 H_0 : 각 집단이 변수 B 의 범주에 대해 동일한 비율을 가진다.
 H_1 : 각 집단이 변수 B 의 범주에 대해 동일한 비율을 가지지 않는다.
- 교차분석을 할 때에는 데이터가 교차표 형태로 주어졌을 때 빈도임을 감안하여 반드시 빈도 변수에 가중값을 부여해 주어야 한다.

- 기대도수가 0을 포함하거나 5보다 작은 셀이 전체의 20%를 초과하면 더 이상 카이제곱분포를 가정할 수 없기 때문에 사용하지 않는다.
- 카이제곱 적합성 검정 : 단일 표본에서 한 변수의 범주값에 따라 기대빈도와 관측빈도 간에 유의미한 차이가 있는지 검정
- χ^2 검정통계량 값은 관측도수와 기대도수의 차이가 크면 커진다. 또한 검정통계량 값이 크면 유의확률은 줄어든다.

< 분산 분석 >

- 분산분석은 집단의 수가 3개 이상인 경우 모평균을 비교 ($\mu_1 = \mu_2 = \mu_3$) 하자 할 때 사용한다. 실험계획법에서 필수적으로 사용된다.
- 분산분석의 오차항에 대한 기본 가정
 - ① 독립성 ② 정규성 ③ 등분산성
- 실험계획법의 기본 원리
 - ① 랜덤화의 원리 : 편의 bias 방지 ② 반복의 원리 : 오차항의 자유도 증가로 오차 분산 감소
 - ③ 불작화의 원리 : 실험 정도 증가 ④ 교락의 원리 : 오차의 교호작용과 교락
 - ⑤ 적교화의 원리 : 검정력 향상
- 요인, 인자 (factor) : 집단을 나타내는 변수, 회귀분석의 독립변수를 분산분석에서는 요인이라 한다.
- 모수인자 : 기술적으로 미리 정해진 수준이 사용되며, 각 수준이 기술적인 의미를 가지고 있는 인자. (온도, 압력, 방식)
- 변량인자 : 수준의 선택이 랜덤으로 이루어지며 각 수준이 기술적인 의미를 가지고 있지 못한 인자. (날씨, 사람)
- 수준 (level) 또는 처리 (treatment) : 요인이 갖는 값
- 교호작용 : 2 인자 이상의 특정한 인자수준의 조합에서 일어나는 효과

- 일원배치 분산분석 (One-way ANOVA) $H_0: \mu_1 = \mu_2 = \dots = \mu_k$, $H_1: \text{모든 } \mu_i \text{ 가 같은건 아니다.}$

일원배치 분산분석은 집단을 나타내는 변수인 인자의 수가 1개인 경우이며 완전확률화 계획법 (CRD)라고도 한다.

- 모집단에 대한 가정 : 각 모집단의 분포가 정규분포를 따르며 서로 독립이고, 모분산은 모두 동일하다.
- 검정통계량 F 값과 기각치 $F_{\alpha; k-1, n-k}$ 를 비교하여 $F = \frac{MSA}{MSE} > F_{\alpha; k-1, n-k}$ 이면 귀무가설 기각, F 의 $p\text{-value} < \alpha$ 이면 기각.
- 이원배치 분산분석 (Two-way ANOVA)

집단을 나타내는 변수인 인자의 수가 2개인 경우이다. 반복 유무에 따라 반복이 있는 이원배치 분산분석과 반복이 있는 분석으로 구분한다.

반복이 있는 이원배치 분산분석에서는 2 인자 이상의 특정한 인자수준의 조합에서 일어나는 효과인 교호작용을 검출 가능.

- 반복이 없는 이원배치 분산분석 : 두 개의 인자가 모두 모두인자일 경우 사용하며, 교호작용 효과는 검출할 수 없다.

① 모수 모형 : 두 개의 인자가 모두 모두인자인 경우

② 혼합 모형 : 하나의 인자는 모두인자이고 다른하나는 변량인자인 경우

$$H_0: \mu_1 = \mu_2 = \dots = \mu_l, \quad \mu_1 = \mu_2 = \dots = \mu_m$$

→ 귀무가설이 기각되었다면 인자수준 간에 모평균 차가 있다는 것을 의미.

2 인자의 수준 조합에서의 모평균의 추정 역시 모든 인자가 유의한 경우 (귀무가설 기각)에 의미가 있다.

- 반복이 있는 이원배치 분산분석 : 반복이 있는 이원배치법 이상에서는 교호작용을 오차항과 구별하여 구할 수 있으므로 주 효과에 대한 검출력이 높아진다. 만약 교호작용이 유의하다면 주 효과에 대한 유의성을 알 수 있으므로 주 효과에 대한 분석은 무의미하다고 할 수 있다.

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

$$\mu_1 = \mu_2 = \dots = \mu_m$$

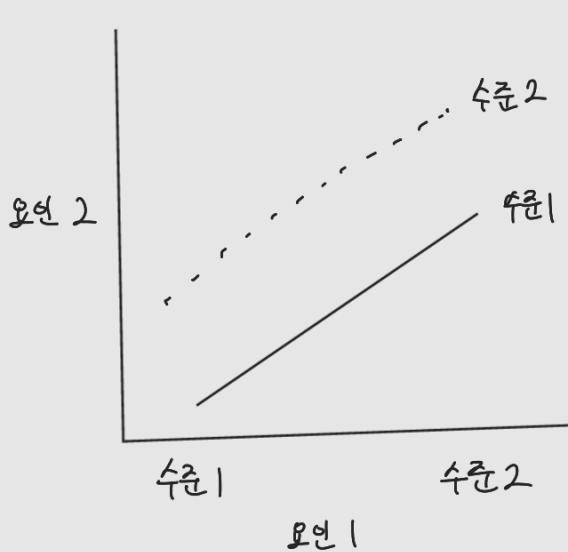
$$\mu_1 = \mu_2 = \dots = \mu_m$$

◦ 교호작용 효과

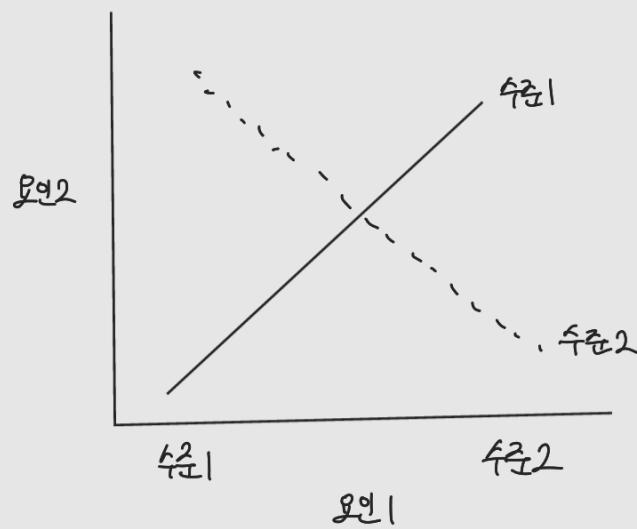
: 두 요인 이상의 특정한 요인 수준의 조합에서 일어나는 효과를 교호작용이라고 한다. 교호작용 효과는 반복이 있는 이원배치법 이상에서 검출할 수 있으며 교호작용 $A \times B$ 가 유의한 경우 요인 A 와 B 의 각 수준의 모평균을 추정하는 것은 일반적으로 무의미하다.

그러므로 실험 결과의 해석은 교호작용을 중심으로 이루어진다. 두 요인에 대한 수준 수가 각각 2 일 때 교호작용 효과가 있는 경우와 없는 경우를 그래프로 나타내면 다음과 같다.

교호작용 효과가 없는 경우



교호작용 효과가 있는 경우



◦ 혼합 모형의 분산 분석

① 난괴법 : 난괴법은 A 인자는 모두 인자이고 B 인자는 변량인자인 경우의 반복이 없는 이원배치 분산분석으로 확률화 블럭계획법 (Randomized Block Design) 이라고도 한다.

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

난괴법에서는 인자 B가 변량인자이므로 인자 B의 모평균 추정은 의미가 없고 수준간의 산포 (σ_B^2)를 추정하는 것만이 의미가 있으며 모수인자 A의 모평균 추정만 의미가 있다.

변량인자인 B (블록)에 대한 분산비 $F = \frac{V_B}{V_E}$ 는 유의성 검정을 위한 것이 아니라,

이원배치 분산분석 설계의 효율성을 평가하는데 쓰인다.

② 반복이 있는 이원배치 혼합모형 : A는 모수인자이고 B는 변량인자인 경우의 반복이 있는 이원배치 분산분석을 혼합모형이라 한다. $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

반복이 있는 이원배치 혼합모형에서는 변량인자 B에 대해서 모평균이나 두 인자의 수준 조합 $A_i B_j$ 에서의 모평균 추정은 의미가 없다. 단지 수준간의 산포 (σ_B^2) 와 교호작용의 σ_{AB}^2 을 추정하는 것만이 의미가 있으며 모수인자 A의 모평균 추정만이 의미가 있다.

F 검정통계량은 $F = \frac{V_A}{V_{A \times B}}$ 을 사용한다.

◦ 일원 배치 분산분석의 분산 분석 표

| 요인 | 제곱합 (SS) | 자유도 | 평균제곱 (MS) | F 값 | F_α |
|----|--|-------|-------------------------|-----------------------|------------------------|
| 처리 | $SSA = \sum_i \sum_j (\bar{x}_{ij} - \bar{\bar{x}})^2$ | $k-1$ | $MSA = \frac{SSA}{k-1}$ | $F = \frac{MSA}{MSE}$ | $F_{\alpha; k-1, n-k}$ |
| 잔차 | $SSE = \sum_i \sum_j (x_{ij} - \bar{x}_{ij})^2$ | $n-k$ | $MSE = \frac{SSE}{n-k}$ | | |
| 합계 | $SST = \sum_i \sum_j (x_{ij} - \bar{\bar{x}})^2$ | $n-1$ | | | |

◦ 일원배치 모수 모형의 구조식

$$x_{ij} = \mu + \alpha_i + e_{ij} = \mu + (\mu_i - \mu) + e_{ij}, e_{ij} \sim iid N(0, \sigma_e^2)$$

단, $\alpha_i = \mu_i - \mu$, $\sum_{i=1}^k \alpha_i = 0$, $i = 1, 2, \dots, k$, $j = 1, 2, \dots, n$

◦ 변동의 분해

인자 A의 수준 수가 k개이고, 각 수준마다 반복이 r회인 경우, 각각의 데이터 x_{ij} 와 총평균 $\bar{\bar{x}}$ 와의 차 $(x_{ij} - \bar{\bar{x}})$ 를 다음과 같이 분해할 수 있다.

$$(x_{ij} - \bar{\bar{x}}) = (x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{\bar{x}})$$

즉, 총평차 $(x_{ij} - \bar{x})$ 는 잔차 $(x_{ij} - \bar{x}_{i\cdot})$ 와 각 수준이 가지고 있는 효과의 크기 $(\bar{x}_{i\cdot} - \bar{x})$ 로 나누낼 수 있다. 양변을 제곱하고, 잔차의 합 $\sum_{j=1}^r (x_{ij} - \bar{x}_{i\cdot}) = 0$ 이므로 총평차 제곱합은 다음과 같이 표현할 수 있다.

$$\sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x})^2 = \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x}_{i\cdot})^2 + \sum_{i=1}^k \sum_{j=1}^r (\bar{x}_{i\cdot} - \bar{x})^2$$

총제곱합 또는 총변동

Total Sum of Squares

SST

잔차제곱합 또는 그룹내변동

Within Sum of Squares

SSW

처리변동 또는 그룹간변동

Between Sum of Squares

SSB

• 반복이 없는 이원배치 분산분석 (인자의 수준은 l, m)

모수 모형 : $X_{ij} = \mu + a_i + b_j + e_{ij}$, $e_{ij} \sim iid N(0, \sigma^2_E)$

단, $\sum_{i=1}^l a_i = 0$, $\sum_{j=1}^m b_j = 0$, $i = 1, 2, \dots, l$, $j = 1, 2, \dots, m$

반복이 없는 이원배치 분산분석표

| 요인 | 제곱합(SS) | 자유도(ϕ) | 평균제곱(V) | F | F_α |
|----|---------|---------------|--------------------------|-------------------|----------------------------|
| A | S_A | $l-1$ | $\frac{S_A}{l-1}$ | $\frac{V_A}{V_E}$ | $F_\alpha; \phi_A, \phi_E$ |
| B | S_B | $m-1$ | $\frac{S_B}{m-1}$ | $\frac{V_B}{V_E}$ | $F_\alpha; \phi_B, \phi_E$ |
| E | S_E | $(l-1)(m-1)$ | $\frac{S_E}{(l-1)(m-1)}$ | | |
| T | S_T | $lm - 1$ | | | |

