AI+X선도인재양성기초프로젝트

5. 연속형 데이터 분석

Acknowledgement

한서경, 확률과 확률 분포, 의학통계론, 서울대학교

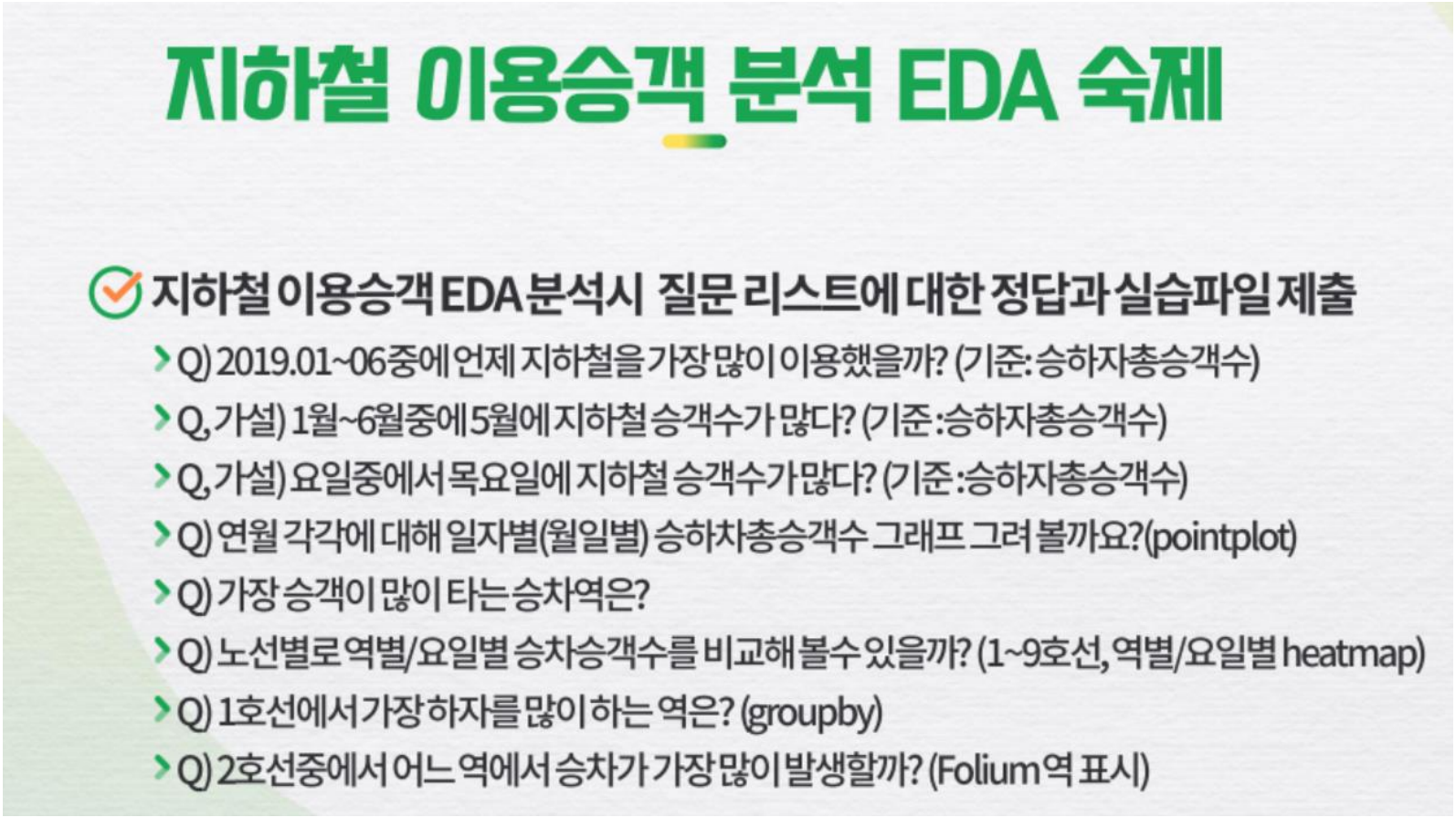
박병주, 평균치의 통계적 분석, 의학통계론, 서울대학교

Heenam Yoon

Department of
Human-Centered Artificial Intelligence

E-mail) h-yoon@smu.ac.kr
Room) 0112

Review

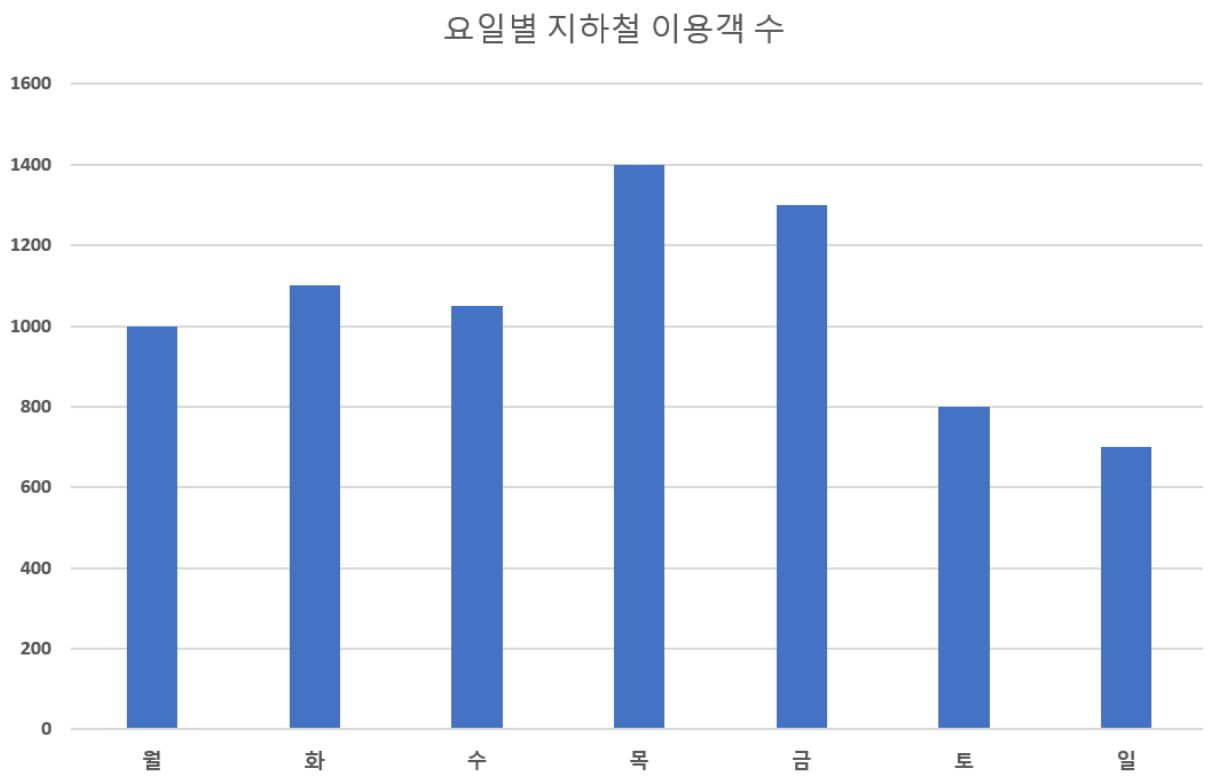


예. 2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

OO 승차역은 다른 역 대비 승하차 총 승객수가 많은가?

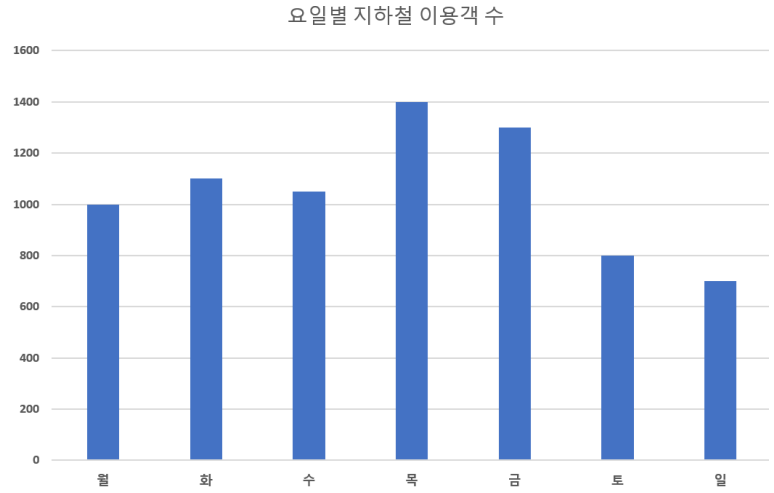
2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

- 2019년 1월 ~ 6월까지 매주 요일별 이용객수를 계산
- 요일별 평균 값으로 다음과 같이 도식화



2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가?

- 예를들어 3월 둘째주 목요일 어떤 행사가 있어 지하철 이용자 수가 10,000명 이었음
- 그 외 목요일은 대략 1,000명 +/- a 수준으로 크게 다르지 않음
- 목요일이 다른 요일 보다 이용객의 수가 많다고 할 수 있을까?



1월 1째주	1000
1월 2째주	950
…	
3월 2째주	10,000
3월 3째주	900
…	
6월 4째주	970
평균	1400

- 목요일이 다른 요일과 비교하여 승객수가 차이가 있다는 것을 어떻게 평가할까?

통계 분석

예. 목요일의 승객수는 다른 요일의 승객수와
통계적으로 유의미한 차이를 보인다
($p < 0.05$)

OO 승차역은 다른 역 대비 승하차 총 승객수가 많은가?

- 지하철 이용 관련 정보를 이용하여 점포 위치 (승하차역)를 추천하고자 한다
- 승하차 승객수, 환승역 여부, 요일별 이용자 수 등 다양한 정보 활용 가능

- 승하차 승객수가 중요한 정보인가?
- 고매출/저매출 점포가 위치한 지하철 역의 승객수는 다를까?

정보 → 

→ 점포 위치 추천

- 고매출/저매출 점포가 위치한 지하철 역의 승객수가 차이가 있다는 것은 어떻게 평가할 것인가?

통계 분석

예. 고매출/저매출 점포가 위치한 지하철 역의 승객수는 통계적으로 유의미한 차이를 보인다 ($p < 0.05$)

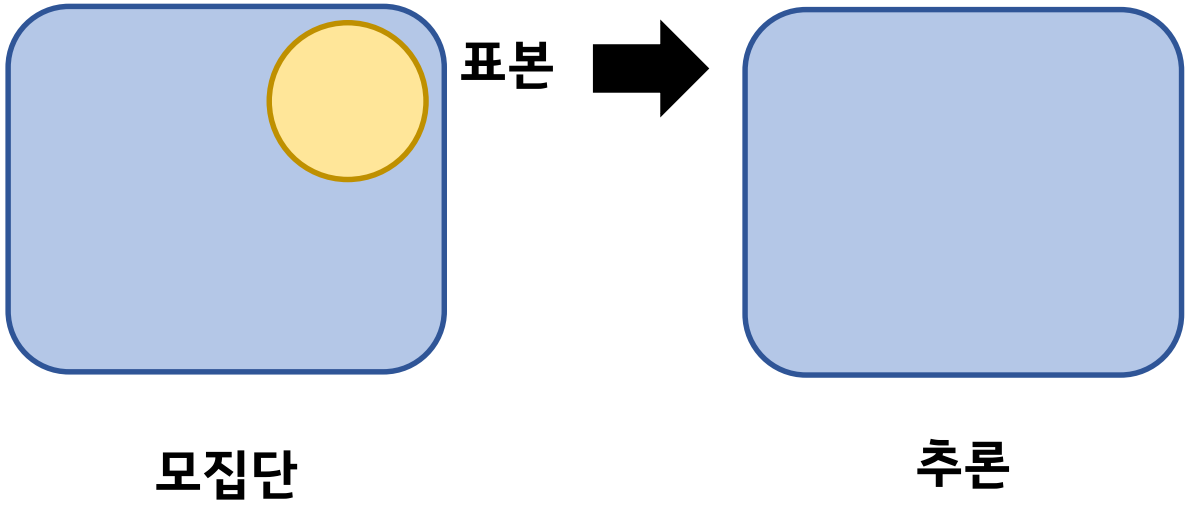
- **자료의 형태, 자료의 수, 비교 방법 등에 따라 통계분석법은 다르다**

모집단과 표본

- **모집단 (Population)**
  - 특정한 특성을 가진 모든 관찰 대상 집단

- **표본 (Sample)**
  - 모집단의 일부. 관찰된 부분 집합

통계적 추론



통계 기초

통계적 추론 예.

데이터 (자료)의 분류

- **데이터 분류**
  - 연속형: 심박수, 성적(0-100), …
  - 범주형
    - 명칭척도: 혈액형(A, B, AB, O), 성별(남/여), …
    - 순위척도: 성적(A, B, C, …), 부서평가 (상, 중, 하, …)

확률 분포: 정규 분포

- 통계학에서 가장 중요한 분포
- 대칭적인 종모양의 분포
- 많은 측정치는 대부분 대략적으로 정규분포를 따름
- 평균과 분산으로 곡선이 정의됨
- 분산이 클수록 납작하고 평평한 분포를 이룸

확률 분포: 표준 정규 분포 (Z)

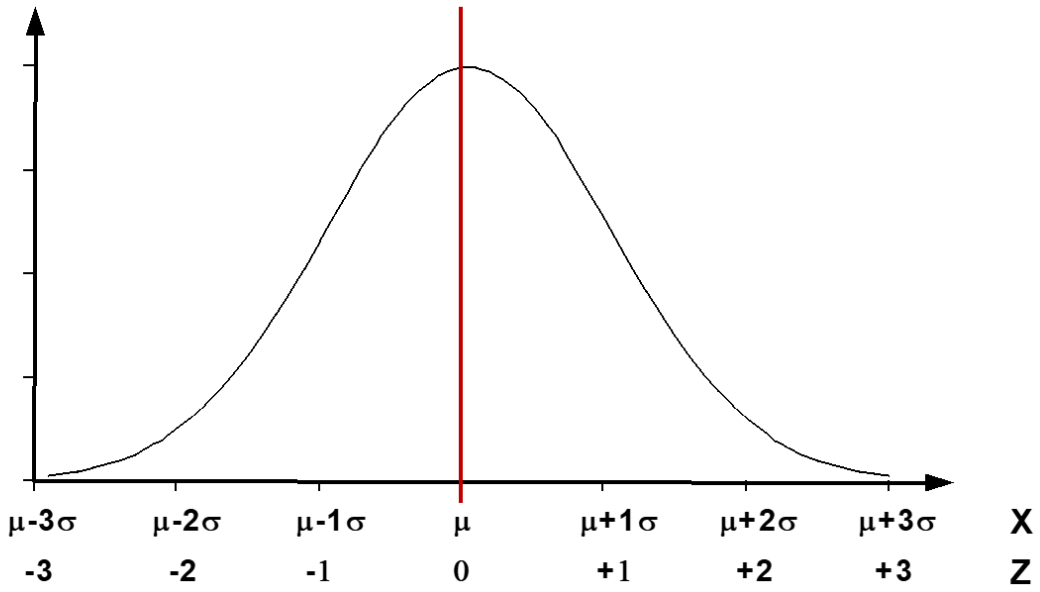
- 무수히 많은 정규 분포가 정의 가능
- 표준화된 정규분포(표준 정규 분포)를 기준으로 확률을 설명하는 것이 쉬움
- 평균이 0 분산이 1인 정규 분포

$$Z = \frac{X - \mu}{\sigma}$$

통계 기초

확률 분포: 표준 정규 분포 (Z)

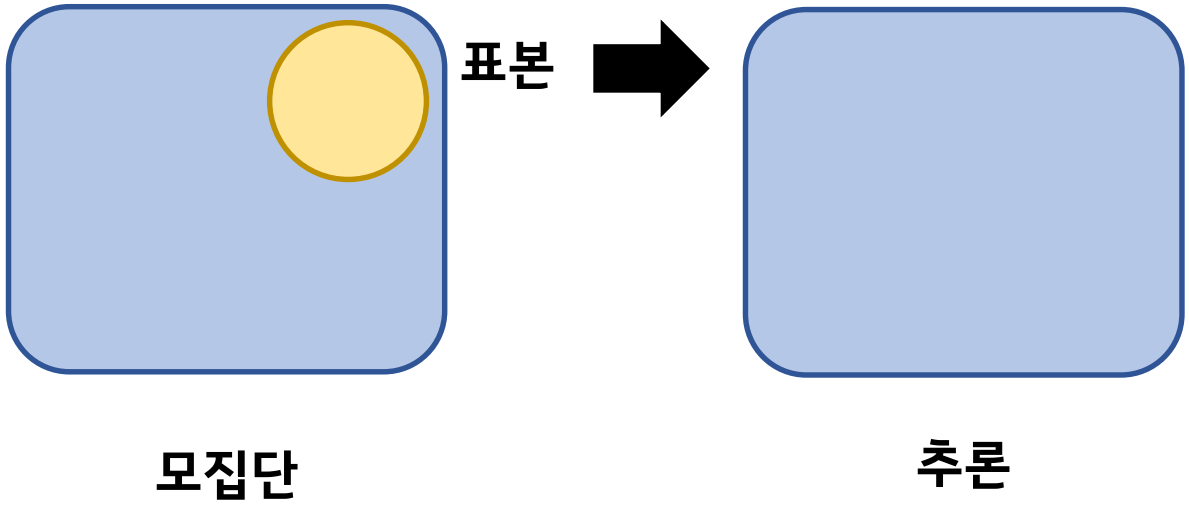
$$Z = \frac{X - \mu}{\sigma}$$



신뢰 구간 (Confidence Interval, CI)

- 모수가 일정 확률로 존재할 것이라고 추정되는 구간
  - 모수: 모집단을 설명하여 요약한 수치 (예. 평균)
- 여기서 일정 확률을 신뢰 수준(Confidence Level)이라고 함

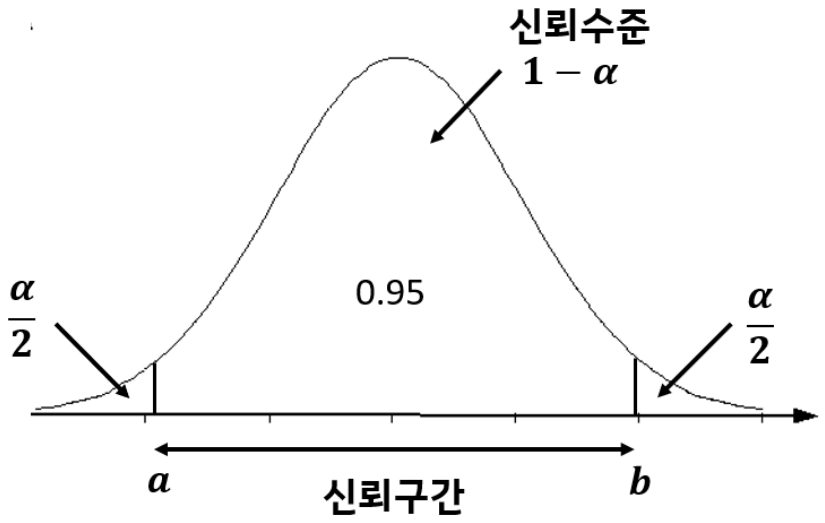
$$a \leq \mu \leq b$$ (95% 신뢰수준)



신뢰 구간 (Confidence Interval, CI)

- 모수가 신뢰구간에 포함되지 않을 확률을 보통 α로 표현

$$P(a \leq \mu \leq b) = 1 - \alpha$$



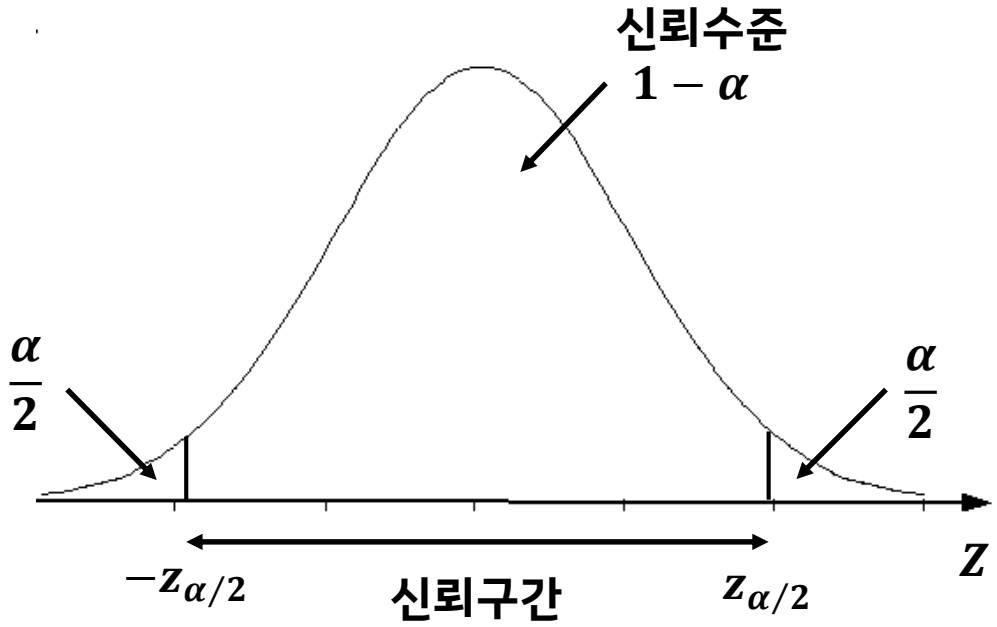
신뢰 구간 (Confidence Interval, CI)

- 모집단의 분산을 아는 경우 통계량은 다음과 같음

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

$$P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$P(a \leq \mu \leq b) = 1 - \alpha$$



신뢰 구간 (Confidence Interval, CI)

$$P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

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

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

신뢰 구간 (Confidence Interval, CI)

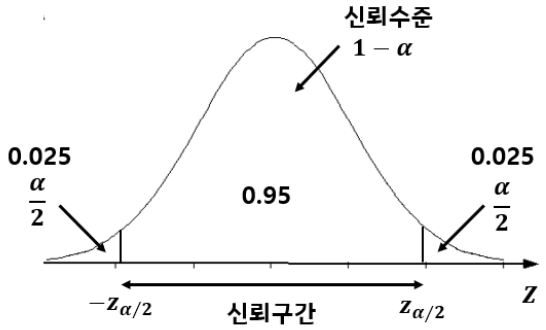
- 예시. 목요일의 지하철 승객수 조사. 모집단 표준편차 500명. 이때 2020년~2021년 100주 동안 목요일 지하철 승객수를 조사했더니, 평균이 1,500명. 목요일의 지하철 승객수에 대한 95% 신뢰구간을 추정하라

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

통계 기초

신뢰 구간 (Confidence Interval, CI)

- Z table

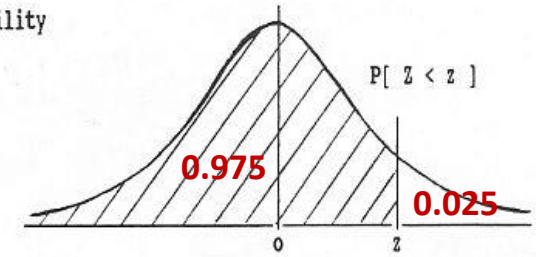


STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $P[\,Z < z\,] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}Z^2)\, dZ$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936

https://www.resourcesystemsconsulting.com/2010/05/20/z-table/

신뢰 구간 (Confidence Interval, CI)

- 예시. 목요일의 지하철 승객수 조사. 모집단 표준편차 500명. 이때 2020년~2021년 100주 동안 목요일 지하철 승객수를 조사했더니, 평균이 1,500명. 목요일의 지하철 승객수에 대한 95% 신뢰구간을 추정하라

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

$$1500 - 1.96\frac{500}{\sqrt{100}} \leq \mu \leq 1500 + 1.96\frac{500}{\sqrt{100}}$$

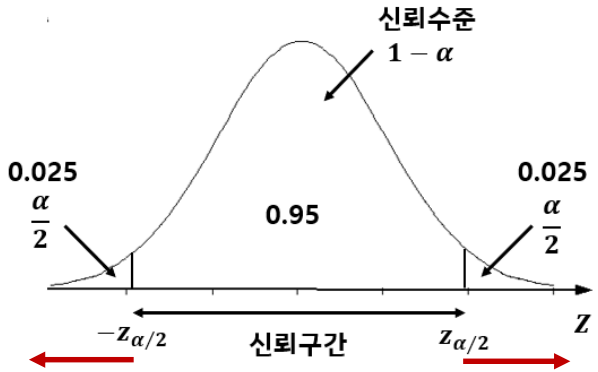
$$1402 \leq \mu \leq 1598$$

해석: 목요일의 지하철 승객수는 1402명에서 1598명 범위 내에 95%의 신뢰수준으로 존재한다.

신뢰 구간 (Confidence Interval, CI)

- 예시. 목요일의 지하철 승객수 조사. 모집단 표준편차 500명. 이때 2020년~2021년 100주 동안 목요일 지하철 승객수를 조사했더니, 평균이 1,500명. 목요일의 지하철 승객수에 대한 95% 신뢰구간을 추정하라

$$1402 \leq \mu \leq 1598$$



$$1000 \leq \mu \leq 2000 \ ?$$

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

$$1492 \leq \mu \leq 1508 \ ?$$

통계 기초

- 자료의 형태, 자료의 수, 비교 방법 등에 따라 통계분석법은 다르다

- 가설을 확인한다
- 데이터의 형태를 확인한다
- 자료의 수를 확인한다
- 통계 검정법을 정한다
- 가설의 발생 가능성을 평가한다

통계 기초

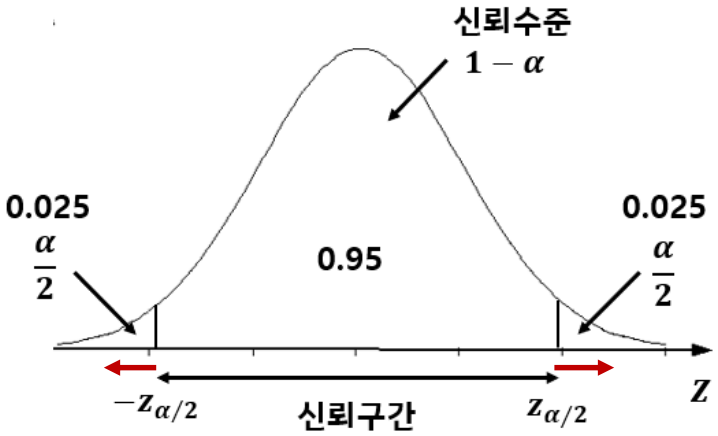
귀무가설(H_0)과 대립가설(H_1)

- **대립가설(Alternative Hypothsis)=연구가설=유지가설**
  - 보통 연구자가 실험을 통해 알고자 하는 가설
  - 목요일 승객수는 1400명과 차이가 있을 것이다
- **귀무가설=영가설(Null Hypothesis)**
  - 대립가설의 반대 가설
  - 목요일 승객수는 1400명과 차이가 없을 것이다

- **통계 분석은 귀무가설이 틀림을 입증하는 것**
  - 목요일 승객수가 1400명과 같을 확률이 5% 미만이라면, 목요일 승객수가 1400명과 차이를 보인다고 말할 수 있다
  - $p < 0.05$

귀무가설(H_0)과 대립가설(H_1)

- H_1에서 $\bar{x}$의 출현확률은 언제나 높을 것으로 예상
  - 연구자가 기대하고 있는 상황이기 때문
  - 예. 목요일 승객수는 1400명과 차이가 있을 것이다
- H_0에서의 $\bar{x}$ 출현 확률(=p-value)도 같이 높으면 H_1이라고 주장할 수 없음
- H_0에서의 $\bar{x}$ 출현 확률이 낮다면, $\bar{x}$이 H_1이라고 주장할 수 있음

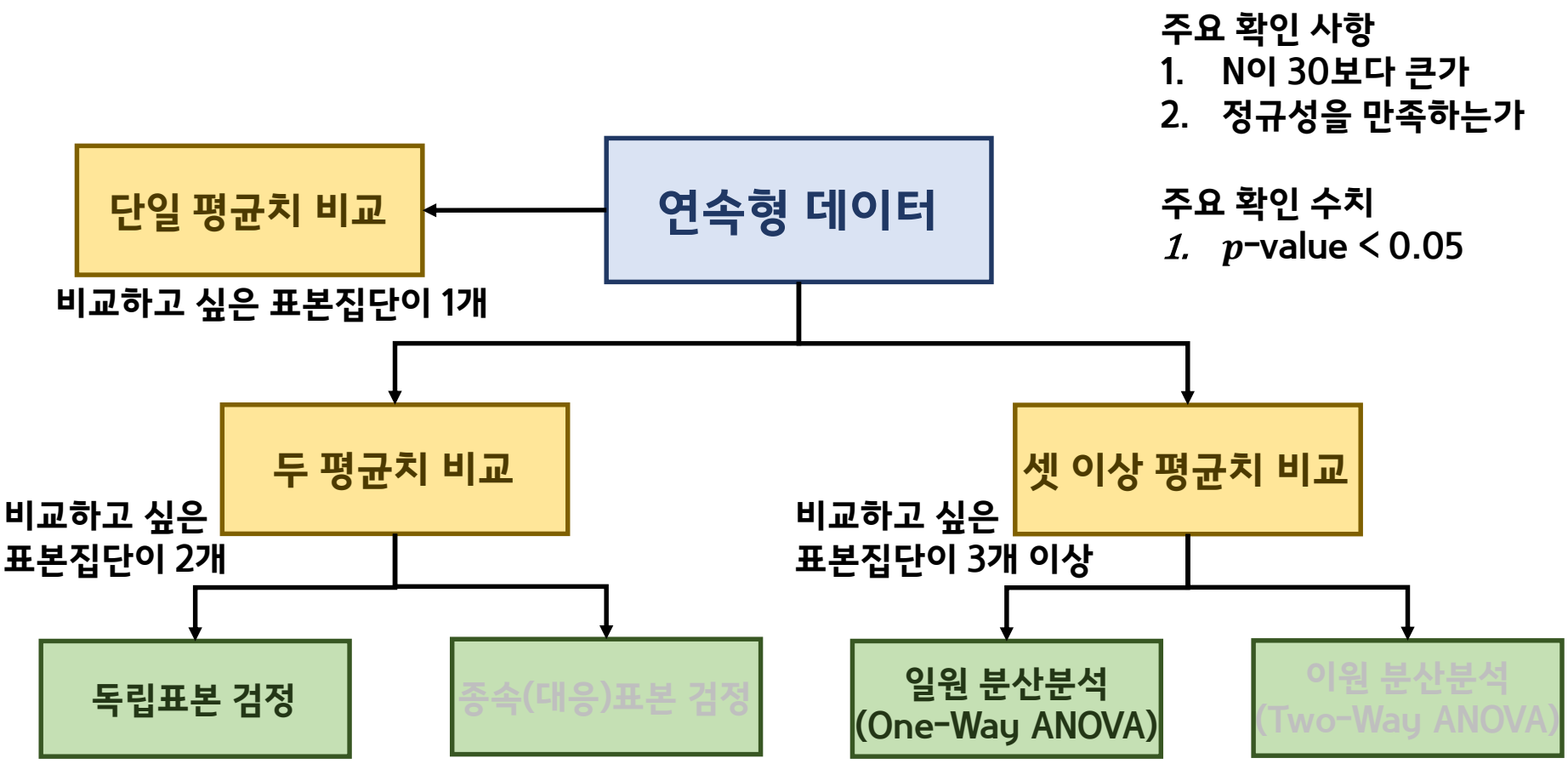


$z > 1.96$보다 크면
H_0기각, H_1채택

데이터 (자료)의 분류

- **데이터 분류**
  - 연속형: 심박수, 성적(0-100), …
  - 범주형
    - 명칭척도: 혈액형(A, B, AB, O), 성별(남/여), …
    - 순위척도: 성적(A, B, C, …), 부서평가 (상, 중, 하, …)

연속형 데이터 분석 요약



연속형 데이터 분석

단일 평균치 비교

- 2020년 ~ 2021년 100주의 목요일 지하철 이용객 수의 평균은 1503.5명이었고, 표준편차는 600.5 bpm이었다. 이는 유의수준 5%에서 1400명과 차이가 있다고 판단 할 수 있겠는가?

- 귀무가설: 1400명과 차이가 없다
- 대립가설: 1400명과 차이가 있다

연속형 데이터 분석

단일 평균치 비교

- $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 로 산출되는 출현확률이 0.05 ($z = \pm 1.96$) 이상이 되는 μ의 범위

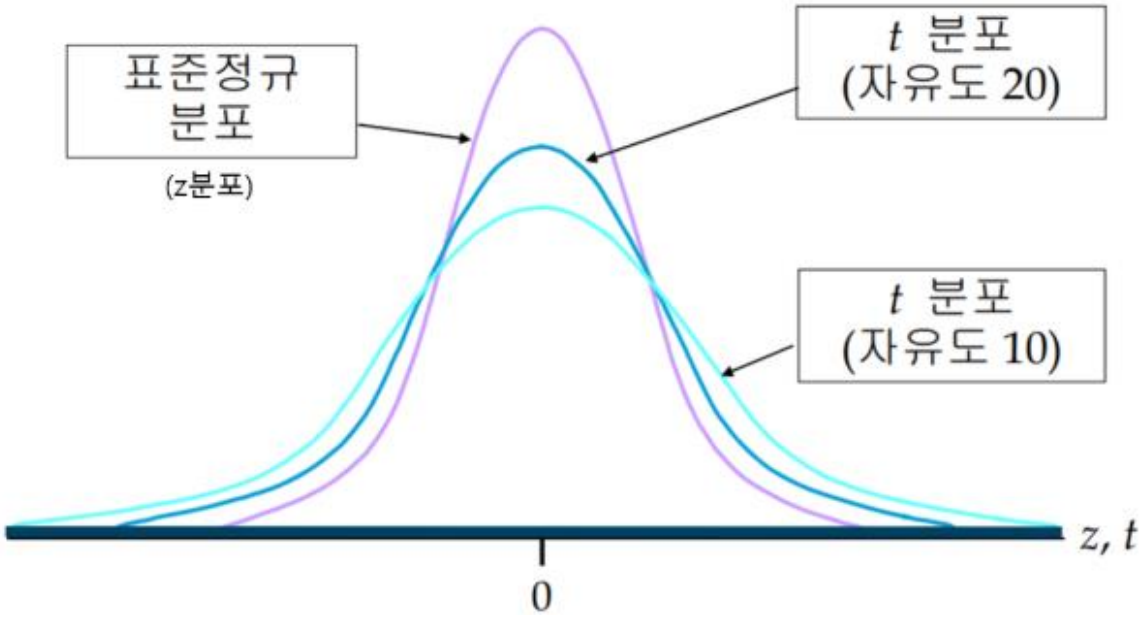
$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = 1.96$$

- σ (모집단의 표준편차)를 알 수 없는 경우가 대부분
- $n \geq 30$이면, 정규분포를 가정하고 t-검정 진행
- $n < 30$이면, 정규성 검정을 진행
  - 만족 시 위와 같은 방법
  - 아닌 경우 비모수 검정

연속형 데이터 분석

t-분포

- 모표준편차 정보 없이 표본의 정보를 이용하여 분포를 추정
- 정규분포보다 꼬리가 두껍고, 표본이 증가할수록 정규분포에 가까워짐



연속형 데이터 분석

t-테이블

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

http://www.ttable.org/

연속형 데이터 분석

단일평균치비교

- 2020년 ~ 2021년 100주의 목요일 지하철 이용객 수의 평균은 1503.5명 이었고, 표준편차는 600.5 bpm이었다. 이는 유의수준 5%에서 1400명과 차이가 있다고 판단 할 수 있겠는가?

- $t = \frac{(1503.5-1400)}{600.5/\sqrt{100}} = 1.724$
- 자유도 99 (df=n-1)에서 $\alpha = 0.05$에 해당하는 t값 1.984 보다 작음
- 면적이 0.05보다 크다는 의미 = 출현확률이 5% 이상
- p > 0.05로 귀무가설을 기각할수 없고, 목요일 지하철 이용객 수의 평균은 1400명과 차이를 보인다고 할 수 없다

연속형 데이터 분석

두 평균치 비교

- 2020년 ~ 2021년 100주의 목요일 지하철 이용객 수는 1503.5±600.5명이었고, 월요일 지하철 이용객 수는 1003.5±200.5명이었다. 두 집단의 차이를 인정할 수 있겠는가 (유의수준 0.05)?

연속형 데이터 분석

두 평균치 비교

- $(\mu_{thr}, \sigma_{thr})$에서 n_{thr}개씩 뽑아서 나온 $\bar{x}_{thr}$와 $(\mu_{mon}, \sigma_{mon})$에서 n_{mon}개씩 뽑아서 나온 $\bar{x}_{mon}$의 차이 $(\bar{x}_{thr} - \bar{x}_{mon})$들의 분포는

  $(\mu_{thr} - \mu_{mon}, \sqrt{\frac{\sigma_{thr}^2}{n_{thr}} + \frac{\sigma_{mon}^2}{n_{mon}}})$인 정규분포를 따름

$$\frac{(\bar{x}_{thr} - \bar{x}_{mon}) - (\mu_{thr} - \mu_{mon})}{\sqrt{\sigma_{thr}^2/n_{thr} + \sigma_{mon}^2/n_{mon}}}$$

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

연속형 데이터 분석

두 평균치 비교

- 2020년 ~ 2021년 100주의 목요일 지하철 이용객 수는 1503.5±600.5명이었고, 월요일 지하철 이용객 수는 1003.5±200.5명이었다. 두 집단의 차이를 인정할 수 있겠는가 (유의수준 0.05)?

$$\frac{(\bar{x}_{thr}-\bar{x}_{mon})-(\mu_{thr}-\mu_{mon})}{\sqrt{\sigma_{thr}^2/n_{thr}+\sigma_{mon}^2/n_{mon}}} \rightarrow 0$$

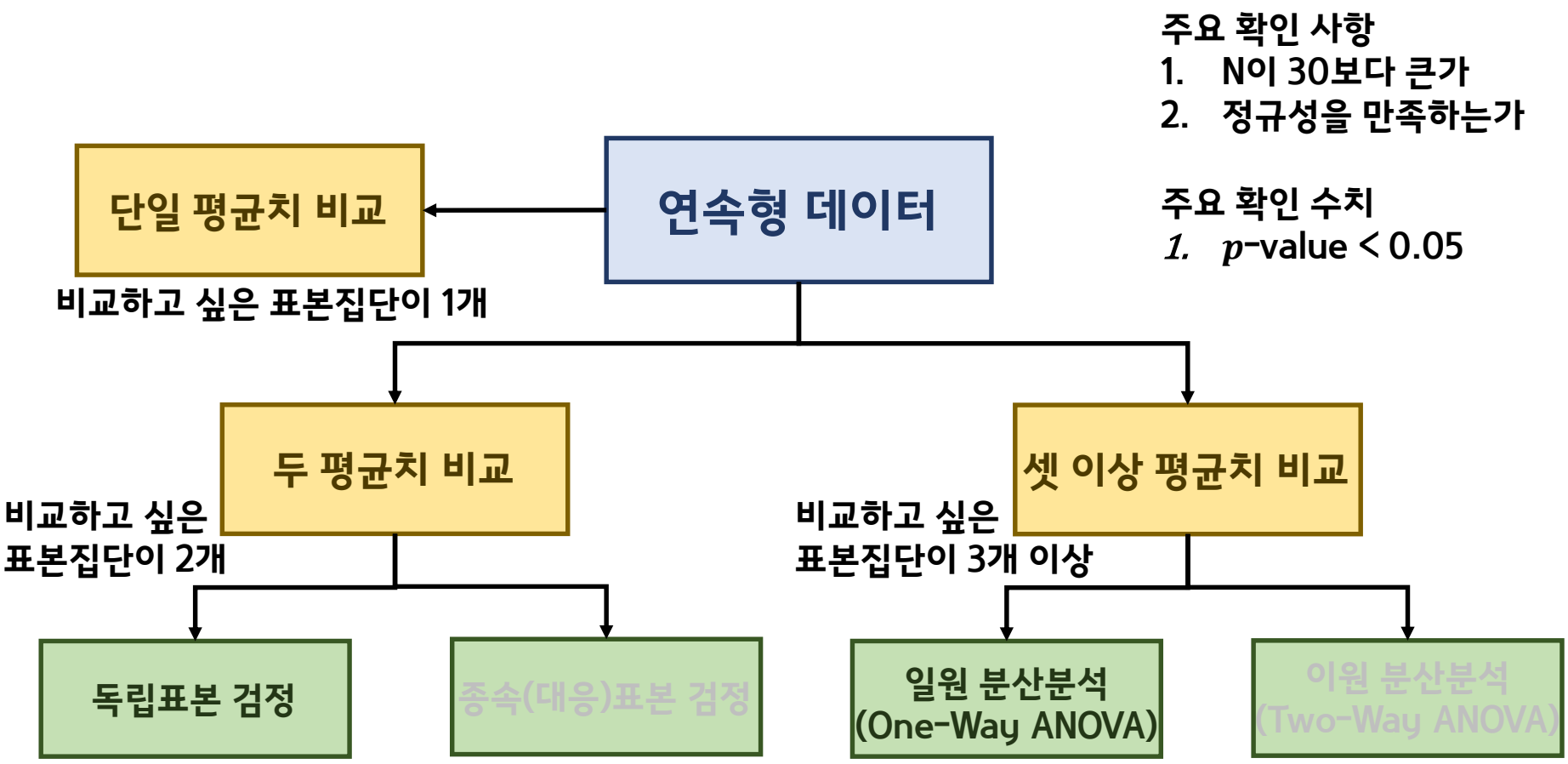
- 귀무가설 (H_0): $\mu_{thr}-\mu_{mon}=0$, 두 평균 간에는 차이가 없다

- $\frac{(\bar{x}_{thr}-\bar{x}_{mon})-(\mu_{thr}-\mu_{mon})}{\sqrt{\sigma_{thr}^2/n_{thr}+\sigma_{mon}^2/n_{mon}}}=\frac{1503.5-1003.5}{\sqrt{600.5^2/100+200.5^2/100}}=7.8978$

- $\alpha=0.05$에 해당하는 값 1.984보다 크므로 $p<0.05$
- 즉, 두 집단의 차이가 인정됨

  = 목요일과 월요일 지하철 승객수에 차이가 있다

연속형 데이터 분석 요약



연속형 데이터 분석

일원 분산분석: 셋 이상의 비교

- 2020년 ~ 2021년 100주의 월요일, 목요일, 토요일 지하철 이용객 수는 차이가 있는가 (유의수준 0.05)?

- 그룹이 3개

3개 이상의 평균치 비교: 일원 분산분석

배경

표본1: $n_m, \bar{x}_m, s_m$ **표본2:** $n_t, \bar{x}_t, s_t$ **표본3:** $n_s, \bar{x}_s, s_s$

- **3개의 표본에 대해 비교를 한다면?**
  - 2개씩 비교, 3번의 T / Z-test가 필요
  - 유의수준 95% 0.95, α = 0.05
  - 올바른 결정을 할 확률 0.95^3
  - $1-0.95^3 = 14.26\%$. $\alpha = 0.1426$?
  - 귀무가설이 참임에도 귀무가설을 기각할 가능성이 14.26% !
  - 동시에 비교해야 함

[표 7-1] 제1종 오류와 제2종 오류

검정 결과 \ 실제 상황	H_0 가 참	H_1 이 참
H_0 를 채택	올바른 결정	제2종 오류
H_1 을 채택	제1종 오류 α	올바른 결정

3개 이상의 평균치 비교: 일원 분산분석

- **분산분석**
  - 세 개 이상의 정규 모집단의 모평균에 요인별로 차이가 있는지를 검정하는 방법

  - 요인 수가 1개인 일원분산분석 (one-way ANOVA)
  - 요인 수가 2개인 이원분산분석 (two-way ANOVA)

3개 이상의 평균치 비교: 일원 분산분석

예

- **예. 일원분산분석**
  - 20대, 30대, 40대, 50대의 스마트폰 하루 사용시간
  - 각 그룹에서 100명씩 스마트폰 하루 사용시간 조사
  - 20대, 30대, 40대, 50대 스마트폰 하루 사용시간에 차이가 있는가?

	20대	30대	40대	50대
1	10	8	6	1
2	7	6	5	5
3	3	2	5	2
4	8	5	3	3
5	12	7	4	2
6	…	…	…	…

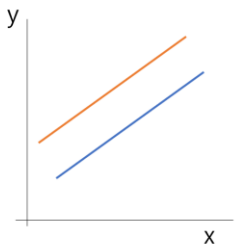
3개 이상의 평균치 비교: 이원 분산분석

예

- **예. 이원분산분석**
  - 남, 녀 / 20대, 30대, 40대, 50대의 심박수
  - 각 그룹에서 남, 녀 100명씩 심박수 측정
  - 남, 녀 / 20대, 30대, 40대, 50대 심박수 에 차이가 있는가?

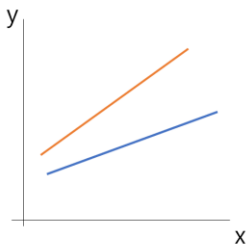
	20대	30대	40대	50대
남	65	63	61	59
녀	70	69	68	67

교호작용 없음



* 기울기가 같고,
* 절편이 다르다.

교호작용 존재



* 기울기가 다르고,
* 절편도 다르다.

3개 이상의 평균치 비교: 일원 분산분석

- 우리가 알고 싶은 것은 집단 간의 통계적 유의한 차이
- 집단은 각각 독립적 = 다른 모집단 (예. 월요일, 목요일, 토요일 승객수)

- 예를 들어 집단이 3개 (편의상 1, 2, 3)
- 1, 2가 차이가 있을 수 있고, 1, 3이 차이가 있을 수 있고, 2, 3이 차이가 있을 수 있고, 1, 2, 3, 모두가 차이를 보일 수도 있음

- 각각 비교는 앞서 언급한 바와 같이 1종 오류가 증가
- 한번에 비교 필요

- 2번의 스텝으로 나눔

3개 이상의 평균치 비교: 일원 분산분석

- 유의한 차이를 보이는 집단이 있는지 (= 다른 모집단에서 출현한 것이 있는지)
- 있다면, 어떤 집단 간에 차이를 보이는지 (사후분석)

3개 이상의 평균치 비교: 일원 분산분석

- 유의한 차이를 보이는 집단이 있는지
  - 관측한 집단에서 유의한 차이를 보이는 집단이 없다

    = H0 = 차이가 없다 = 같은 모집단에서 나온 것이다

    H1 = 다른 모집단에서 출현한 집단이 적어도 한 개 이상 존재한다

    예를 들어 집단이 3개 (편의상 1, 2, 3)

    1, 2가 차이가 있을 수 있고, 1, 3이 차이가 있을 수 있고, 2, 3이 차이가 있을 수 있고, 1, 2, 3, 모두가 차이를 보일 수도 있음

3개 이상의 평균치 비교: 일원 분산분석

표본1: $n_1, \bar{x}_1, s_1$ **표본2:** $n_2, \bar{x}_2, s_2$ **표본3:** $n_3, \bar{x}_3, s_3$

- **귀무가설 (H_0): 같은 모집단에서 출현**

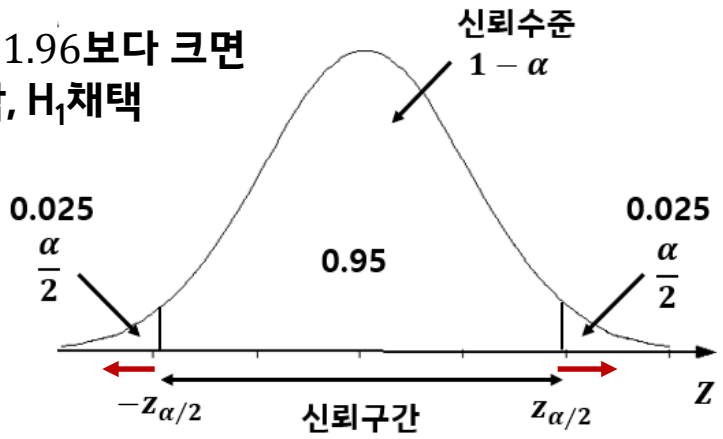
- **Between Variance**
  - 각 그룹의 평균들이 모평균 μ에 대해 가지는 편차 (분산)
- **Within Variance**
  - 개별 자료들이 그룹평균에 대해 가지는 편차(분산)

3개 이상의 평균치 비교: 일원 분산분석

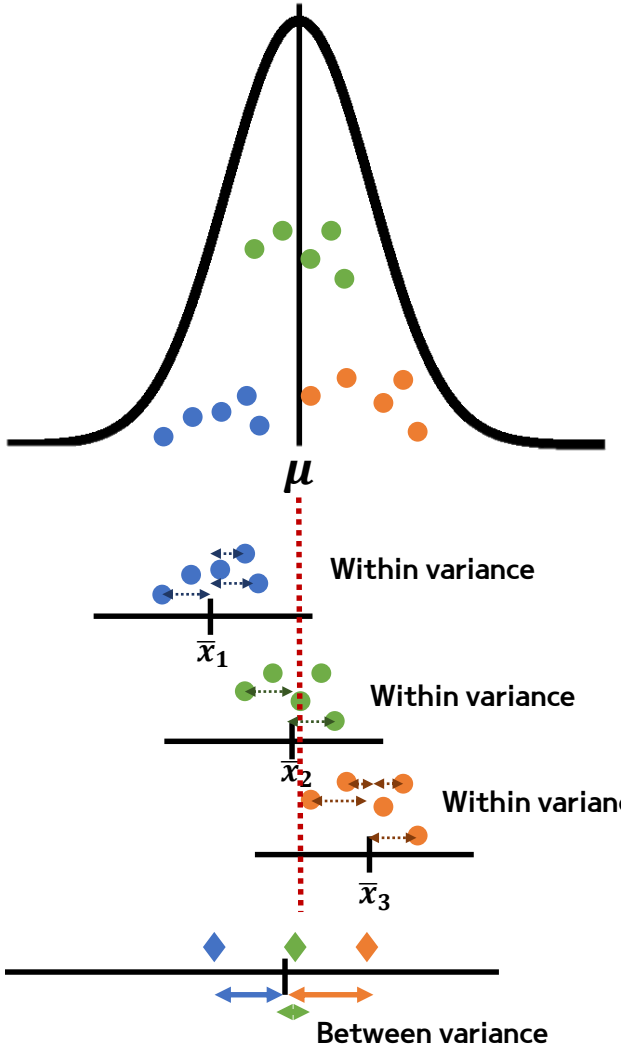
- (복습) 가설 검정에서 하는 일
  - 귀무가설 세우기
  - 관측 값 계산
  - 관측 값이 기각 영역에 놓이는지 확인
  - 기각 영역에 놓이면 귀무가설 기각 & 대립가설 채택

$|z_0| > 1.96$**보다 크면**
H_0기각, H_1채택

$|z_0|$**가 클수록 좋음!**



3개 이상의 평균치 비교: 일원 분산분석



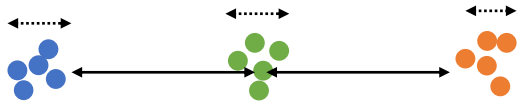
$n_1 = 5$

$n_2 = 5$

$n_3 = 5$

귀무가설 (H_0): 같은 모집단에서 출현

- Between Variance
  - 각 그룹의 평균들이 모평균 μ에 대해 가지는 편차 (분산)
- Within Variance
  - 개별 자료들이 그룹평균에 대해 가지는 편차(분산)



같은 모집단에서 출현(H_0) 했다고 말할 수 있을까?

Between variance ↑ Within variance ↓

위 상황을 잘 표현할 수 있는 식?

$$\frac{Between\ Variance}{Within\ Variance}$$ ↑ ↓ ↑

3개 이상의 평균치 비교: 일원 분산분석

$$\frac{Between\ Variance}{Within\ Variance}$$

분산의 비

$$F^{\uparrow} = \frac{Between\ Variance^{\uparrow}}{Within\ Variance_{\downarrow}}$$

분산을 분석한다

요인이 1개: one-way

ANalysis Of VAriance: ANOVA

3개 이상의 평균치 비교: 일원 분산분석

- **가설검정 절차**

집단1 (y1)	집단2 (y2)	집단3 (y3)
y11	y21	y31
y12	y22	y32
y13	y23	y33
…	…	…

❶ 처리 효과가 존재하지 않는다는 것은 k 개 집단의 모집단이 동일하다는 의미이므로, 일원분산분석의 가설은 다음과 같다.

H0 = 차이가 없다 = 같은 모집단에서 나온 것이다

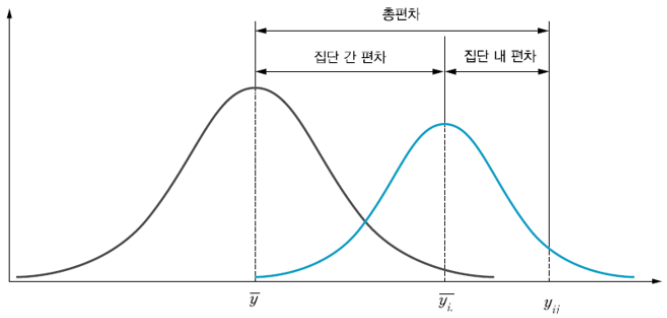
귀무가설 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$(모든 집단의 모평균이 같다)

대립가설 $H_1 : H_0$가 아니다(모든 집단의 모평균이 모두 다 같지는 않다)

H1 = 다른 모집단에서 출현한 집단이 적어도 한 개 이상 존재한다

❷ 위 가설을 검정하기 위해서 총편차를 다음과 같이 분해한다.

$$\underbrace{y_{ij} - \bar{y}}_{\text{총편차}} = \underbrace{(\bar{y_{i.}} - \bar{y})}_{\text{집단 간 편차}} + \underbrace{(y_{ij} - \bar{y_{i.}})}_{\text{집단 내 편차}}$$



[그림 8-1] 분산분석에서의 편차

3개 이상의 평균치 비교: 일원 분산분석

- **가설검정 절차**

앞 식의 양변을 제곱하고 모든 i, j에 대하여 합하면 다음 결과를 얻는다.

집단간 편차의 제곱 합 **집단내 편차의 제곱 합**

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{\text{총제곱합}} = \underbrace{\sum_{i=1}^{k} n_i(\overline{y_{i.}}-\bar{y})^2}_{\text{처리제곱합}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2}_{\text{오차제곱합}}$$

SSt **SSE**

총제곱합(total sum of squares)은 편차 제곱의 총합으로, SST로 표기한다. **처리제곱합**(treatment sum of squares)은 처리(집단) 간 편차의 제곱합으로, **급간제곱합**(between sum of squares)이라고도 하며 SSt로 표기한다. **오차제곱합**(error sum of squares)은 집단 내 편차의 제곱합으로, **급내제곱합**(within sum of squares)이라고도 하며 SSE로 표기한다.

3개 이상의 평균치 비교: 일원 분산분석

- **가설검정 절차**

앞 식의 양변을 제곱하고 모든 i, j에 대하여 합하면 다음 결과를 얻는다.

집단간 편차의 제곱 합 **집단내 편차의 제곱 합**

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{\text{총제곱합}} = \underbrace{\sum_{i=1}^{k} n_i(\overline{y_{i.}}-\bar{y})^2}_{\text{처리제곱합}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_{i.}})^2}_{\text{오차제곱합}}$$

SSt **SSE**

표본으로부터 계산한 처리제곱합과 오차제곱합의 크기로 다음과 같이 귀무가설에 대한 판단을 내릴 수 있다.

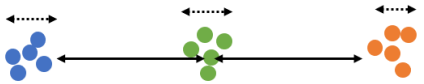
- 처리제곱합이 크고 오차제곱합이 작다면, 귀무가설이 틀릴 가능성이 크다.
- 처리제곱합이 작고 오차제곱합이 크다면, 귀무가설이 옳을 가능성이 크다.

처리제곱합과 오차제곱합의 크기를 비교함으로써 다음과 같은 검정통계량 F를 만들 수 있다. 이때 제곱합을 자유도로 나누면 어떤 미지의 분산의 추정량이 되며, 이를 **평균제곱(mean square)**이라고 한다.

$$F = \frac{SSt/(k-1)}{SSE/(n-k)} = \frac{MSt}{MSE} \sim \mathrm{F}(k-1, n-k), \quad n = \sum_{i=1}^{k} n_i$$

귀무가설 (H_0): 같은 모집단에서 출현

- Between Variance
  - 각 그룹의 평균들이 모평균 μ에 대해 가지는 편차 (분산)
- Within Variance
  - 개별 자료들이 그룹평균에 대해 가지는 편차(분산)



같은 모집단에서 출현(H_0) 했다고 말할 수 있을까?

Between variance ↑ **Within variance ↓**

위 상황을 잘 표현할 수 있는 식?

$$\frac{Between\ Variance\ \uparrow}{Within\ Variance\ \downarrow} \uparrow$$

3개 이상의 평균치 비교: 일원 분산분석

- **가설검정 절차**

표본으로부터 계산한 처리제곱합과 오차제곱합의 크기로 다음과 같이 귀무가설에 대한 판단을 내릴 수 있다.

- 처리제곱합이 크고 오차제곱합이 작다면, 귀무가설이 틀릴 가능성이 크다.
- 처리제곱합이 작고 오차제곱합이 크다면, 귀무가설이 옳을 가능성이 크다.

처리제곱합과 오차제곱합의 크기를 비교함으로써 다음과 같은 검정통계량 F 를 만들 수 있다. 이때 제곱합을 자유도로 나누면 어떤 미지의 분산의 추정량이 되며, 이를 **평균제곱(mean square)**이라고 한다.

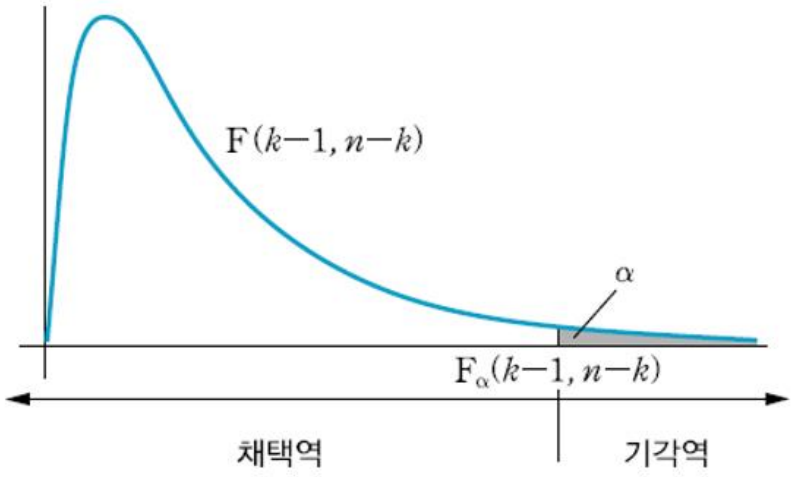
Between variance

$$F = \frac{SSt/(k-1)}{SSE/(n-k)} = \frac{MSt}{MSE} \sim \mathrm{F}(k-1, n-k), \quad n = \sum_{i=1}^{k} n_i$$

Within variance

3개 이상의 평균치 비교: 일원 분산분석

- 가설검정 절차



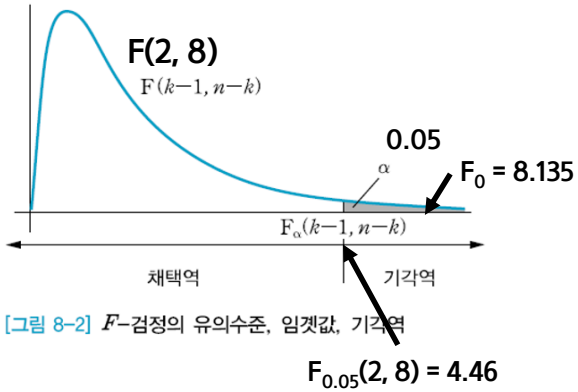
[그림 8-2] F-검정의 유의수준, 임곗값, 기각역

3개 이상의 평균치 비교: 일원 분산분석

예제 8-2

다음은 금연을 돕기 위한 세 가지 프로그램의 효율성을 비교하기 위해 조사한 각 프로그램별 지원자의 효과점수이다. 프로그램별 효과가 다르다고 할 수 있는지를 유의수준 5%로 검정하라.

프로그램	프로그램 1	프로그램 2	프로그램 3
효과점수	13.5	10.3	19.0
	18.0	11.8	18.4
	14.1	13.1	15.3
	17.8	–	17.3
평균	15.85	11.73	17.5
전체 평균	15.33		



[그림 8-2] F-검정의 유의수준, 임곗값, 기각역

앞 식의 양변을 제곱하고 모든 i, j에 대하여 합하면 다음 결과를 얻는다.

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}_{\text{총제곱합}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{y_i}-\bar{y})^2}_{\text{처리제곱합}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y_i})^2}_{\text{오차제곱합}}$$

집단간 편차의 제곱 합 = SSt

집단내 편차의 제곱 합 = SSE

SSt: 58.725
SSE: 28.877
k-1: 3 - 1 = 2
n-k: 11 - 3 = 8

$F_{0.05}(2, 8) = 4.46$

$F_0 = 8.135$

$$F = \frac{SSt/(k-1)}{SSE/(n-k)} = \frac{MSt}{MSE} \sim F(k-1, n-k), \quad n = \sum_{i=1}^{k} n_i$$

다른 모집단에서 출현한 표본이 적어도 1개 이상 있다
= 금연 효율성이 다른 프로그램이 적어도 1개 이상 존재한다

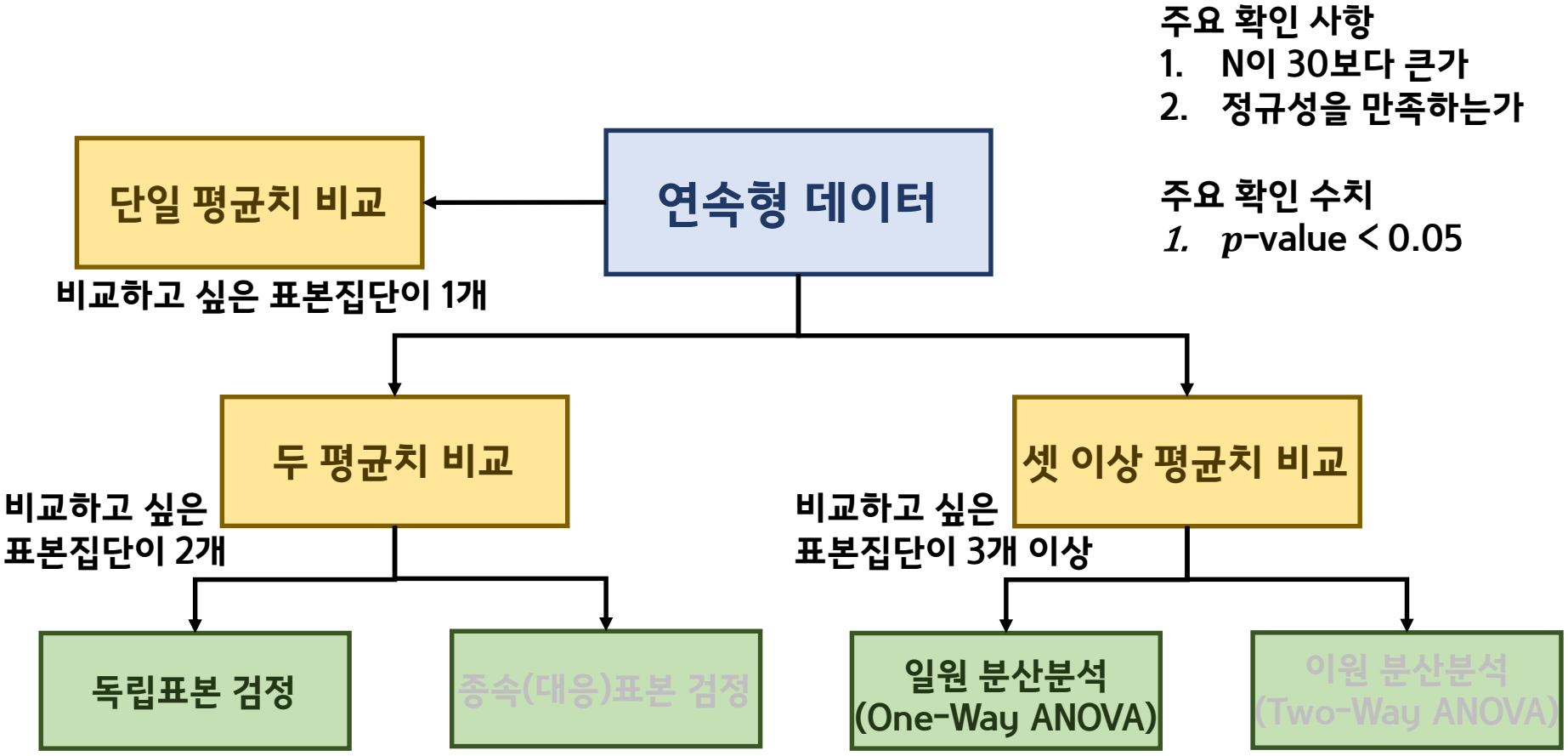
3개 이상의 평균치 비교: 일원 분산분석

해석

- F-test 하여 non-significant: 종결
  - $\bar{x}_1 = \bar{x}_2 = \bar{x}_3$

- 유의하다고 나온 경우 ($p < 0.05$)
  - 개별 평균들을 비교하여 추론
  - Tukey test, Duncan test

Summary



Summary

- 2020년 ~ 2021년 100주의 목요일 지하철 이용객 수의 평균은 1503.5명이었고, 표준편차는 600.5 bpm이었다. 이는 유의수준 5%에서 1400명과 차이가 있다고 판단 할 수 있겠는가?

- 2020년 ~ 2021년 100주의 목요일 지하철 이용객 수는 1503.5±600.5명이었고, 월요일 지하철 이용객 수는 1003.5±200.5명이었다. 두 집단의 차이를 인정할 수 있겠는가 (유의수준 0.05)?

- 2020년 ~ 2021년 100주의 월요일, 목요일, 토요일 지하철 이용객 수는 차이가 있는가 (유의수준 0.05)?

Practice

- 2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가 (차이가 있는가)?

- 데이터 형태: 연속형 데이터
- 자료의 수: 1월 ~ 6월, 4주 * 6개월 = 24개

- N >= 30
  - 목 vs. the others:
  - 월 vs. 화 … vs. 일:

만족

- N < 30
- 정규성 검정

불만족

- 비모수검정
  - 목 vs. the others:
  - 월 vs. 화 … vs. 일:

Practice

정규성 검정

- 자료의 정규성을 평가하는 방법
- 다양한 검정 방법이 있음 (Kolmogorov-Smirnov, Shapiro-Wilk, …)
- p-value 로 정규성 여부 판단
  - $p < 0.05$: 정규분포를 따르지 않는다
  - $p > 0.05$: 정규분포를 따른다

- 두 그룹간 유의한 차이가 있다 ($p < 0.05$)
  - 귀무가설: 두 그룹간 유의한 차이가 없다
  - 귀무가설의 가능성이 5%미만이므로 ($p < 0.05$), 귀무가설을 기각하고, 대립가설을 채택했었음

Practice

정규성 검정

- 정규성 검정은 어떤 정규 분포와 우리의 자료의 분포가 차이를 보이는지 평가하는 것
  - 귀무가설: 어떤 정규 분포와 우리 자료의 정규분포는 유의한 차이가 없다
  - 대립가설: 어떤 정규 분포와 우리 자료의 정규분포는 유의한 차이가 있다
- 귀무가설의 가능성을 평가
- 귀무가설의 가능성이 5% 미만이라면 ($p < 0.05$), 귀무가설이 기각되고, 대립가설을 채택
- p-value 로 정규성 여부 판단
  - $p < 0.05$: 정규분포를 따르지 않는다
  - $p > 0.05$: 정규분포를 따른다

Practice

- 2019년 1월 ~ 6월 중 목요일에 지하철 승객수가 많은가 (차이가 있는가)?

- 데이터 형태: 연속형 데이터
- 자료의 수: 1월 ~ 6월, 4주 * 6개월 = 24개

- N >= 30
  - 목 vs. the others:
  - 월 vs. 화 … vs. 일:

만족

- N < 30
- 정규성 검정

불만족

- 비모수검정
  - 목 vs. the others:
  - 월 vs. 화 … vs. 일:

Thank you.