통계 이해

[기술통계]

기술통계 척도

1. 중심 척도: 평균, 중앙값, 최빈값
    - `평균`: 가장 보편/대표적, outlier에 민감, 수치척도(등간/비율/순서) 대표값
    - `중앙값`: outlier에 둔감
    - `최빈값(Mo)`: 명목척도 대표값
2. 산포 척도: 범위, 분산, 표준편차, 사분위수 범위
    - `범위`: outlier에 민감
    - `분산(` σ^2 `,` s^2 `)`: 편차 제곱의 평균, outlier에 민감, unit 표현 불가

$$s^2 = \frac{\Sigma(y-\bar{y})^2}{n-1} = \frac{SS}{df}$$

    - `표준편차(σ, s)`: 분산의 양의 제곱근, outlier에 민감, unit 표현 가능
    - `사분위수 범위(IQR)`: 중간 50%(Q3-Q1), 범위 척도보다 outlier에 둔감
3. 분포 모양: 도수 분포, 비대칭도(=왜도), 첨도
    - `도수 분포`: 도수, 상대도수, 누적도수
    - `비대칭도(왜도)`: 양수면 오른쪽으로 긴 꼬리(왼쪽 봉우리), 음수면 왼쪽으로 긴 꼬리(오른쪽 봉우리)
    - `첨도`: 양수면 중앙 뾰족(정규분포보다 긴 꼬리), 음수면 중앙 완만(정규분포보다 짧은 꼬리)

데이터의 유형

데이터 유형		분류 (category)	순위 (order)	동일한 간격 (equal interval)	절대영점 (absolute zero)	대표값	통계분석
이산형 데이터	명목척도	○	X	X	X	최빈값, 퍼센트	빈도분석, 비모수통계
	서열척도	○	○	X	X	중위값, 퍼센트	비모수통계
연속형 데이터	등간척도	○	○	○	X	산술평균	모수통계
	비율척도	○	○	○	○	산술평균, 기하평균	

- 연속형 데이터: 등간척도, 비율척도
  - `등간척도`: 가감연산(+, -) 가능, e.g. 온도, 물가지수
  - `비율척도`: 사칙연산(+, -, ×, ÷) 가능, e.g. 거리, 무게, 시간
- 이산형 데이터(범주형 데이터): 명목척도, 순위척도
  - `명목척도`: 사칙연산 불가능, e.g. 성별, 품질, 운동선수 등번호, 종교
  - `순위척도(서열척도)`: 사칙연산 불가능, e.g. 만족도, 학교성적등급, 크기

[확률분포]

- `확률(P)`: 특정사건이 발생할 가능성
- `표본공간(S)`: 실험결과 발생할 수 있는 모든 가능한 결과의 집합
- `확률변수(X)`: 표본공간을 실수에 대응시키는 함수(또는 방법)

> 💡 만약 주사위 2개를 던졌다면, 표본공간은 (1,1), (1,2), ... , (6,6)이고, 확률변수를 '두 눈의 수의 합'으로 정의했을 때 X는 2, 3, ... , 12의 값을 취할 수 있다.

  - `이산확률변수`: 확률변수가 취하는 값이 유한개일 때
  - `연속확률변수`: 확률변수가 구간 내의 임의의 모든 점을 취할 수 있을 때
- `확률함수`: 확률변수에 대하여 정의된 실수를 0과 1 사이의 실수(확률)에 대응시키는 함수
  - `이산확률함수(PMF)`, `연속확률함수(PDF, 확률밀도함수)`, `누적확률함수(CDF)`

- `확률분포`: 모든 가능한 확률변수 값과 그 값이 발생할 확률 값을 도수분포표나 그래프로 나타낸 것
    - `연속확률분포`: 모두 확률밀도함수 존재
        - `정규분포(가우스 분포)`
            - 매개변수: 평균, 표준편차
            - 그래프: 평균을 중심으로 좌우 대칭인 종모양 분포, 곡선 아래 면적 1
            - 기호: X ~ N(μ,σ^2)
            - 용도: 수집된 자료 분포 근사(중심극한정리에 의해 독립적인 확률변수의 평균은 정규분포에 가까워지는 성질이 있기 때문)

            ```
            from scipy import stats
            prob = stats.norm.cdf(x, mu, sigma)
            ```

        - `표준정규분포(z분포)`: 정규분포 밀도함수를 통해 X를 Z로 정규화하여 평균 0, 표준편차 1인 정규분포

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

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

            - 기호: Z ~ N(0,1^2)
            - 용도: z검정
        - `t분포`

            $$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

            - 매개변수: 자유도($df = n - 1$)
            - 그래프: z분포보다 더 넓고, 꼬리 부분이 더 평평
            - 용도: 모평균 추/검정에서 모표준편차를 모를 때 정규분포 대신 사용, 회귀분석에서 개별 회귀계수의 유의성 검정

```
from scipy import stats
prob = stats.t.cdf(t, df)
```

- `카이제곱 분포`: z분포를 제곱하여 합한 것(x≥0), 정규분포를 따르는 모집단에서 크기가 n인 표본을 무작위로 반복하여 추출했을 때, 각 표본에 대해 구한 표본분산은 카이제곱 분포를 따름

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

  - 매개변수: 자유도($df = n - 1$)
  - 기호: X ~ $\chi^2(n-1)$
  - 용도: 모분산 추정, 빈도 기반 분포/형태 적합도 검정, 여러 집단 간 독립성/동질성 검정

```
from scipy import stats
prob = stats.chi2.cdf(chisq, df)
```

- `F분포`: 카이제곱분포 2개를 서로 나눈 값(x≥0), t분포를 제곱하면 F분포, 분산이 같은 두 정규모집단으로부터 크기 n1과 n2의 확률표본을 반복하여 독립 추출한 후 구한 두 표본분산의 비율들의 표본분포
  - 매개변수: $df1, df2$
  - 용도: 두 분포의 분산 비교, ANOVA에서 그룹 내/간 변동으로 여러 개의 평균값을 비교할 때 활용, 회귀분석에서 회귀모형 자체의 유의성 검정

```
from scipy import stats
prob = stats.f.cdf(x=f, dfn=df1, dfd=df2)
```

- `와이블 분포`: 지수분포를 일반화시켜, 여러 다양한 확률분포 형태를 모두 나타낼 수 있도록 고안됨
  - 매개변수: 형상모수(α), 척도모수(β)
  - 그래프: 형상모수 1은 지수분포/2는 라이레히 분포
  - 용도: 수명 분포(설비/부품의 수명 추정, 실패시간, 대기시간)

```
# 어떤 제품의 수명시간 x가 형상모수 2.2, 척도모수 1,200인 와이블 분포를 따름
```

```
# 이 제품이 적어도 1,500 시간 이상 작동할 확률

from scipy import stats
x = 1500
alpha = 2.2
beta = 1200
prob = stats.weibull_min.cdf(x, alpha, scale=beta)

# P(X>=x): 1 - prob = 0.195
```

- 이산확률분포
  - 이항분포: 베르누이 실험을 n번 시행해서 특정한 횟수의 성공/실패가 나타날 확률을 알고 싶을 때, 각 시행마다 성공 확률 p는 항상 일정

$$\Pr(K=k) = f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathrm{E}(X) = np$$

$$\operatorname{Var}(X) = np(1-p).$$

    - 베르누이 시행: 표본공간이 단지 2개의 상호배타적인 원소로 구성된 실험의 시행
    - 베르누이 분포: 이항분포에서 n=1인 특수한 경우
    - 매개변수: p(성공 확률), n(시행 횟수)
    - 그래프: p→0.5, n→∞ 일 때(np≥5, n(1-p)≥5 일 때) 이항분포는 정규분포 곡선에 가까워짐

```
# 도장공정에서 광택도 불량 40%
# 3대 차량을 임의로 선택했을 때 불량대수가 각각 0, 1, 2, 3대가 나올 확률

from scipy import stats
n = 3
for i in range(n+1):
  prob = stats.binom.pmf(k=i, n=n, p=0.4)
  print("P(X={}): {:.3f}".format(i, prob))
```

  - 포아송 분포: 일정한 시/공간에서 발생하는 성공횟수에 대한 이산확률분포
    - 매개변수: m(일정단위당 평균발생 횟수)
    - 용도: 일정 시/공간에서의 사건 발생 확률 예측

```
from scipy import stats

# 1분당 평균 전화가 걸려오는 횟수
mu = 2

# 1분당 3번의 전화가 걸려올 확률
prob_pmf = stats.poisson.pmf(3, mu)

# 1분당 최대 2회 이하의 전화가 걸려올 확률
prob_cdf = stats.poisson.cdf(2, mu)
```

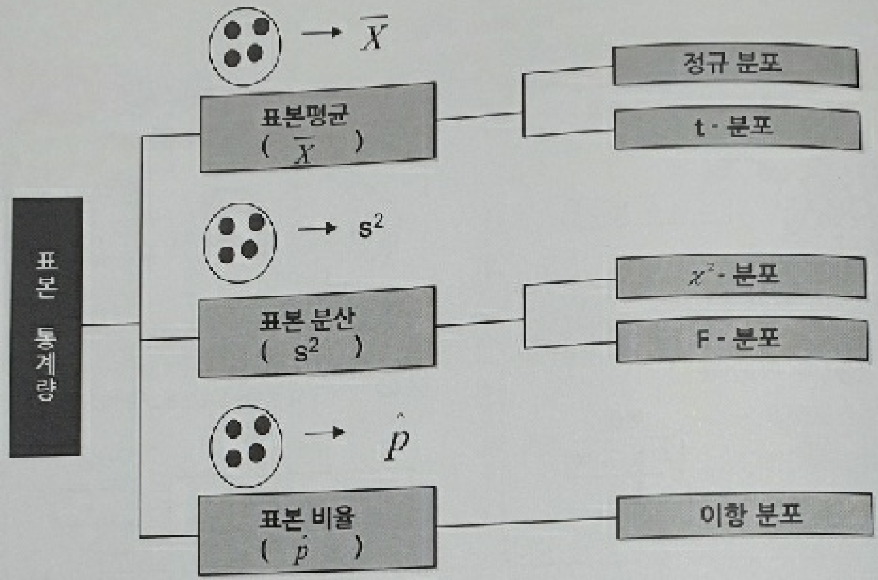
[표본분포]

- `표본추출`: `랜덤 샘플링`, `층별화 샘플링`(층별), `계통적 샘플링`(매 n번째), `서브그룹 샘플링`(매 t시간별 n단위 샘플링)
- 모수 VS. 통계량
  - `모수`: e.g. 모평균, 모표준편차
  - `통계량`: 모수를 추정하기 위해 표본으로부터 계산된 값, e.g. 표본평균, 표본표준편차
- 표본평균의 분포
  - `중심극한 정리`: 모집단의 형태와 상관없이 표본평균의 분포는 빠른 속도로 정규분포에 근접
    - 모집단이 정규분포이면, 표본평균의 분포는 표본크기에 상관없이 언제나 정규분포
    - 모집단이 적어도 대칭형이면, 표본크기는 5~20이면 표본평균의 분포는 정규분포에 가까워짐
    - 최악의 경우, 모집단이 정규분포에서 얼마나 벗어났느냐에 상관없이 표본크기가 최소 30 이상이면 표본평균의 분포는 정규분포에 가까워짐
  - `표준오차(standard error, SE)`: 표본분포의 표준편차
  - `평균의 표준오차(standard error of the mean, SEM)`: 표본평균분포의 표준편차

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

통계량의 확률분포

- 통계량(표본평균, 표본 비율, 표본 분산)



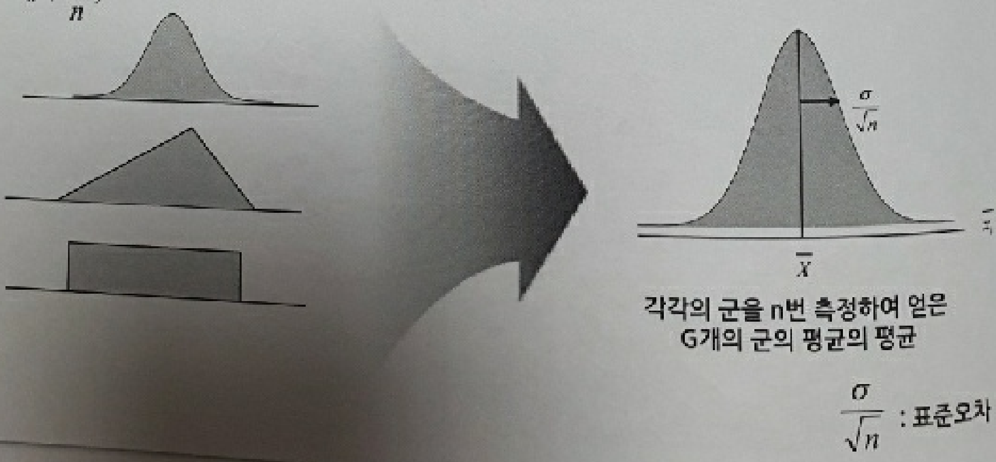
32

표본의 평균 ($\overline{X}$)의 분포_정규분포

- 중심극한 정리(Central Limit Theorem)

: 평균이 μ이고 분산이 σ^2인 임의의 확률분포를 따르는 모집단으로부터 크기 n인 확률표본 $X_1, X_2, \cdots, X_n$을 취했을 때 이 확률표본의 **표본평균($\overline{X}$)의 분포는 표본의 크기 n이 충분히 클 때 대략 다음과 같은 정규분포를 따름**

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

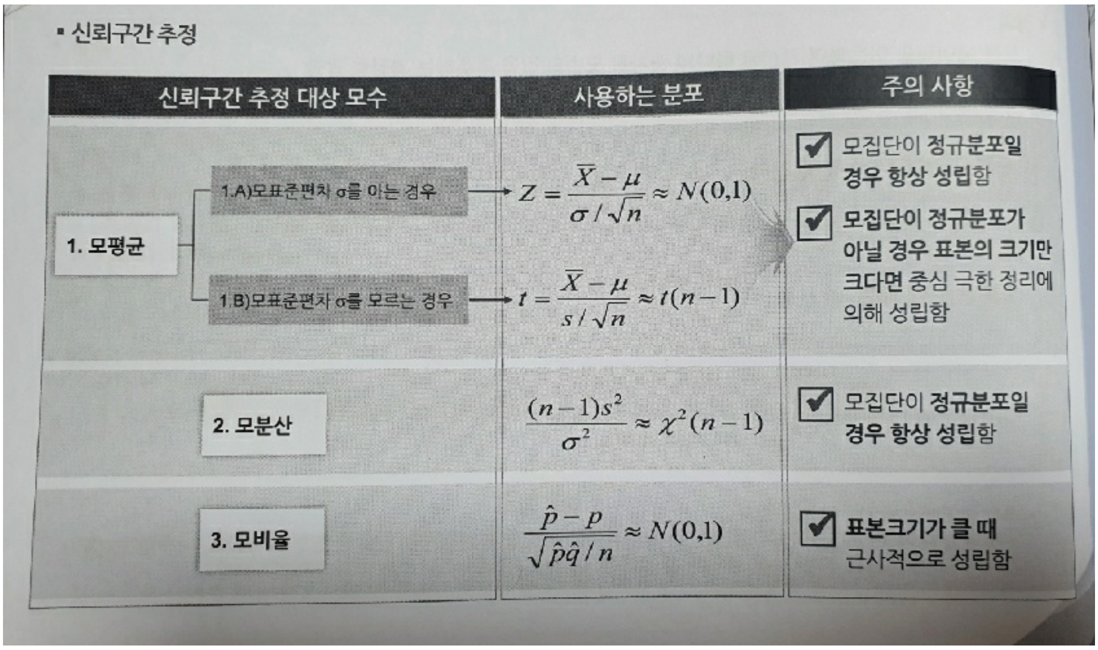


[통계적 추정/검정]

💡 모집단 → (표본추출) → 표본 → (통계량 계산) → 통계량 → (확률분포 선정) → 표본분포 → (가설검정, 신뢰구간 추정) → 모수

점/구간 추정

- `추정` : 모집단에서 추출한 표본에서 얻은 정보를 이용하여 모평균, 모표준편차/모분산, 모비율을 추측하는 것
  - `점추정` : 표본데이터를 이용하여 계산된 하나의 숫자로 모수의 값을 추측하는 과정
    - `추정량` (절차), `추정치` (수치)
  - `구간추정` : 모집단에서 추출한 표본에서 얻은 정보를 이용하여 추정하고자 하는 모수가 존재하리라 예상되는 구간을 정함



▪ 신뢰구간 추정

신뢰구간 추정 대상 모수		사용하는 분포	주의 사항
1. 모평균	1.A)모표준편차 σ를 아는 경우	$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \approx N(0,1)$	모집단이 정규분포일 경우 항상 성립함
	1.B)모표준편차 σ를 모르는 경우	$t = \frac{\bar{X} - \mu}{s / \sqrt{n}} \approx t(n-1)$	모집단이 정규분포가 아닐 경우 표본의 크기만 크다면 중심 극한 정리에 의해 성립함
2. 모분산		$\frac{(n-1)s^2}{\sigma^2} \approx \chi^2(n-1)$	모집단이 정규분포일 경우 항상 성립함
3. 모비율		$\frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q} / n}} \approx N(0,1)$	표본크기가 클 때 근사적으로 성립함

- `신뢰수준` : 추정하고자 하는 모평균이 신뢰구간에 포함될 확률(점추정값 ± 한계오차)
- `모평균 추정`
  - `σ를 아는 경우` : z분포(95%는 1.96, 99%는 2.58)

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

```
# 배추 40통 랜덤추출, 모표준편차 0.397, 모평균 무게에 대한 95% 신뢰구간 추정

from scipy import stats
lower, upper = stats.norm.interval(
    0.95,
    loc=np.mean(df),
    scale=0.397/np.sqrt(40)
)
```

- `σ를 모르는 경우`: t분포

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

```
# 배추 40통 랜덤추출, 모평균 무게에 대한 95% 신뢰구간 추정

from scipy import stats
lower, upper = stats.t.interval(
    0.95,
    df,
    loc=np.mean(df),
    scale=scipy.stats.sem(df)
)
```

- `모분산 추정`: 카이제곱 분포, 모집단이 정규분포를 따를 경우 사용

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

- `모비율 추정`: 이항분포, 모비율 p에 대한 신뢰구간

$$\hat{p} \pm \underbrace{Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{\text{허용오차}}$$

가설 검정

- 절차

    1. 가설 수립: H0/H1, α(유의수준) 결정

        - `H0(귀무가설)`: 기존 사실에 대한 가설, 개선 전 사실, 검정통계량은 귀무가설의 분포에서 나옴, 검정 대상으로 삼는 가설, 차이가 없음(~다), 영향을 주지 않는다는 입장

        - `H1(대립가설)`: 새롭게 확인하고자 하는 사실에 대한 가설, 개선 후 사실, 검정통계량이 귀무가설에서 나왔다고 보기 어려울 경우 대립가설 채택, 귀무가설을 부정하는 가설, 차이가 있다(~가 아니다), 영향은 준다는 입장

        - `α(유의수준)`: 귀무가설을 기각하는 결정이 잘못될 수 있을 최대 가능성

        - `임계값`: 유의수준에서 귀무가설 채택/기각할 때 그 기준이 되는 통계량

    2. 가설 검정: 검정통계량, p-value 계산

        - `p-value`: 귀무가설이 참이라는 가정 하에 표본 데이터가 귀무가설을 지지하는 확률

    3. 검정결과 판단: 검정통계량>임계값 또는 p-value<α 이면 H0 기각

        - `귀무가설을 기각할 수 없다`: 귀무가설이 옳다는 것이 아니라 귀무가설을 기각할 확실한 증거가 없다는 것(즉 귀무가설이 참일 수도 있고 거짓일 수도 있음)

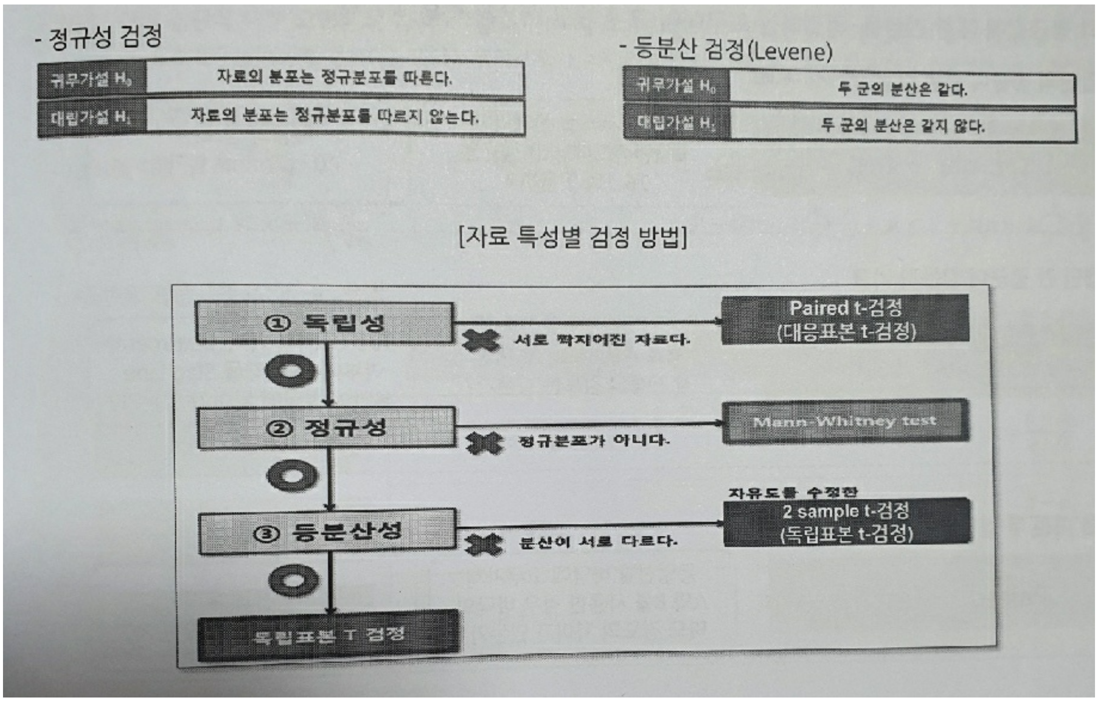
- `정규성 검정`: 확률분포가 정규분포를 따르는지 아닌지 확인하는 것

    - H0: 모집단은 정규분포를 따른다 / H1: 모집단은 정규분포를 따르지 않는다

    - 검정결과: 95% 신뢰수준에서 p-value가 0.05보다 크면 정규, 0.05보다 작으면 비정규

```
from scipy.stats import shapiro
statistic, pvalue = shapiro(series)
```

- 가설검정의 오류
  - `제 1종 오류(α)`: 생산자 위험, 귀무가설을 채택했어야 함에도 불구하고 이를 기각하는 위험, α는 전형적으로 5%로 설정, 초반에 사용자가 결정, 1-α는 신뢰수준
  - `제 2종 오류(β)`: 소비자 위험, 귀무가설을 기각했어야 함에도 불구하고 이를 채택하는 위험, β는 전형적으로 10%로 설정, 다른 모든 값이 동일할 때 α값이 작아지면 β값은 증가, 귀무가설을 기각하는데 많은 증거를 요구하게 되면 제 2종 오류가 일어날 확률이 높아짐, 1-β는 귀무가설이 거짓일 때 이를 기각할 확률 검정력(Power of the test)
- `평균검정`: 평균 차이에 대한 검정
  - H0 채택: 평균 차이는 표본오차에 의한 것(등호 포함) / H0 기각: 평균 차이는 집단의 속성에 의한 것
  - `양측검정`: 차이 유무 / `단측검정`: 가설의 방향이 한 쪽으로 분명한 경우
  - `z검정`: 모집단의 표준편차가 알려져 있을 때
  - `t검정`: 모집단의 표준편차를 모를 때, 모집단이 극단적으로 비정규 분포를 따르지 않는 한 t검정 신뢰구간 추청지는 여전히 타당



  - `1-sample t검정`: 단일 집단의 평균이 기존에 주장하는 평균과 같은지 비교

- 가정: 정규분포, 등분산성

```
# 고객만족도 평균은 76.7, 개선활동을 완료한 후 10개의 고객만족도 데이터를 얻음
# 개선활동이 만족도를 변화시켰는가?
from scipy import stats
t_result = stats.ttest_1samp(df, 76.7)
t, pvalue = t_result.statistic.round(3), t_result.pvalue.round(3)
```

- `2-sample t검정` : 두 집단 간 평균이 같은지 비교
  - 가정: 정규분포, 등분산성

```
from scipy import stats
t_result = stats.ttest_ind(df1, df2)
t, pvalue = t_result.statistic.round(3), t_result.pvalue.round(3)
```

- `Paired t검정` : 쌍을 이룬 두 집단 간 평균이 같은지 비교, 평균 차를 구하여 1-sample t검정과 같은 방법으로 검정
  - 가정: 정규분포

```
from scipy import stats
t_result = stats.ttest_rel(df1, df2)
t, pvalue = t_result.statistic.round(3), t_result.pvalue.round(3)
```

- `비율검정` : 비율 차이에 대한 검정
  - `1 Proportion test` : 한 집단의 비율이 특정 비율과 같은지 비교

```
# A제품을 사용하는 국내 고객은 전체고객 중에 10%
# A제품의 품질개선 결과 전체고객 중 100여 개의 업체를 표본으로 했을 때 15개의 업체가 만족 표현
# 품질개선 결과로 기존보다 전체 고객 중 사용비율의 차이가 있을까?

from statsmodels.stats.proportion import proportions_ztest
stat, pvalue = proportions_ztest(count, n, value)
```

  - `2 Proportion test` : 두 집단의 비율이 같은 지를 검정하는 도구

```
# 동일한 제품을 생산하는 두 공장에서 불량률을 측정한 결과
# 공장1: N1=1000, X1=4
# 공장2: N2=1200, X2=1
# 두 공정의 불량률이 같다고 할 수 있을까?

from statsmodels.stats.proportion import proportions_ztest
count = np.array([4, 1])
```

```
n = np.array([1000, 1200])
stat, pvalue = proportions_ztest(count, n)
```

- `카이제곱 검정` : 관찰된 빈도가 기대되는 빈도와 의미있게 다른지의 여부를 검증하는 검증방법
    - 자료가 빈도로 주어졌을 때, 범주형 자료 분석에 이용
    - `동일성 검정` (차이), `독립성 검정` (관계), `적합도 검정` (기대치)
    - 자유도: 범주의 수-1
    - 카이제곱 검정통계량이 크다는 것은 실측치 대비 기대치의 차이가 크기 때문에 귀무가설 기각 / 카이제곱 검정통계량이 작다는 것은 실측지와 기대치의 차이가 작아 귀무가설을 채택

    ```
    # A별 차이를 본다면, A를 행에 놓을 것!

    from scipy import stats
    chisq, pvalue, df, expected = stats.chi2_contingency(data)
    ```

- `ANOVA(분산분석)` : 집단 간의 평균차이를 검정하기 위해 총변동의 요인을 '수준차이로 설명되는 변동'과 '설명될 수 없는 변동'으로 분해하여 이 두 변동의 비가 통계적으로 유의한지 검정하는 분석방법
    - SS(Total) = SS(Factor) + SS(Error)
        - 총 변동은 2개의 변동으로 분리할 수 있음(집단 내 오차에 의한 것 / 집단 간 수준차에 의한 것)
        - 관심을 가지고 있는 인자가 평균 반응에 대해 영향이 없거나 미미하다면, 이 2개의 추정값은 동일해야 하고, 모든 서브 그룹들은 동일한 모집단에서 온 것으로 결론을 내릴 수 있음
    - `일원분산분석(one-way ANOVA)` : 집단 2개 이상, 독립변수/종속변수 1개
    - `이원분산분석(two-way ANOVA)` : 집단 2개 이상, 독립변수 2개 이상

    ```
    f_result = stats.f_oneway(df['A'], df['B'], df['C'])
    f, pvalue = f_result.statistic.round(3), f_result.pvalue.round(3)
    ```

상관/회귀분석

- `상관분석`

- 공분산: 변수 척도의 단위에 따라 달라짐, 선형의 강도에 대한 정보 제공하지 않기 때문에 특정한 공분산 값이 크고/작은지 결정하기 어려움

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- 상관계수: 변수 척도의 단위에 영향을 받지 않음, 선형적인 관계 강도와 방향을 수치로 표시한 표준화된 지수, 단 두 변수간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아님

$$\text{상관계수} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/(n-1)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}}$$

```
from scipy import stats
corr, pvalue = stats.pearsonr(df1, df2)
```

- 회귀분석: 독립변수가 종속변수에 미치는 영향력의 크기를 측정하여 독립변수의 일정한 값에 대응되는 종속변수의 값을 예측하기 위한 통계적 분석방법
  - 절차: 그래프/상관분석 → 회귀모형 설정 → 다중공선성 검토 → 회귀계수 추정 및 유의성 검증 → 모형 진단 및 잔차 분석
  - 회귀모형 설정: 단순/다중 회귀분석, 선형/비선형 회귀분석

- 회귀계수 추정(최소자승법): 잔차가 최소가 되는 표본회귀식이 구하고자 하는 가장 좋은 회귀식
- 다중공선성: 독립변수들 사이에 상관관계를 갖고 있는 현상, 다중공선성이 존재하면 회귀식 계수의 분산이 매우 커지기 때문에 정확한 모수 추/검정 어려움(두 독립변수의 상관관계가 높으면 종속변수가 동시에 변화할 수 있음), 분산팽창계수(VIF)를 5~10을 넘으면 다중공선성 존재

$$VIF_i = \frac{1}{1 - R_i^2}$$

- 회귀모형 적합도 판정
  1. 결정계수(R^2): 대표적 지표, 회귀식에 의해 설명된 변동이 총변동에서 차지하는 상대적 크기를 나타낸 것
     - 잔차: 실제 관측값 - 회귀모형 예측값
     - 잔차평균제곱: $MSE = \frac{SSE}{n-2}$ (단순회귀모형에서 SSE의 자유도는 n-2)
     - 추정의 표준오차(잔차들의 표준편차): $\sqrt{MSE}$

$$SST = SSE + SSR$$

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

$$R^2 = \frac{SSR}{SST} = (\text{상관계수})^2$$

$\bar{Y}$: 실제 Y값

Y_i: 모집단의 측정치

$\hat{Y}_i$: 표본집단의 예측치

2. `ANOVA` : 분산분석을 통하여 회귀모형의 적합성 검증, 검정통계량으로 잔차평균제곱합과 회귀평균제곱의 비율 이용
    - H0: 독립변수가 설명력이 없다 / H1: 독립변수가 설명력이 있다
3. `잔차분석` : 잔차는 회귀모형을 이루는 여러 가정들의 타당성에 관한 많은 정보를 갖고 있기 때문에 회귀분석에 중요한 역할을 함
    - `정규성` : 표준화된 잔차들을 이용하여 정규확률그림을 그렸을 때 점들이 직선상에 있으면 정규성 만족
    - `등분산성` : 표준화된 잔차들을 이용하여 변수들의 산점도를 그렸을 때 점들이 0을 중심으로 대칭적으로 랜덤하게 나타나고 모두 ±2 범위 내에 있으면 등분산성 만족
    - `독립성` : 시간순서에 따라 잔차 값들을 점찍어 보았을 때 잔차들이 대략 수평대를 형성하면 독립성 만족

- 변수선택: `전진선택법` , `후진제거법` , `단계적 방법` (가장 널리 사용), `모든 가능한 조합의 회귀분석` (모든 독립변수 조합 고려)

```
# 단순선형회귀
from sklearn.linear_model import LinearRegression
import statsmodels.formula.api as smf

model = smf.ols(formula="DV ~ IV", data=df)
result = model.fit()

print(result.summary())
```

스마트

- `데이터 리터러시` : 분석 능력, 창의성, 수학과 통계적 기술, 비즈니스 스킬
- `빅데이터의 특징` (4V)
    - Volume: 수집하는 데이터 양 크게 증가
    - Velocity: 데이터 축적속도가 매우 빠름
    - Variety: 사진, 음성, 문자 등 다양한 종류 데이터
    - Value: 데이터를 활용해 새로운 기회 창출
- `기업에서의 문제해결 방법`

1. `고객중심` : 우리를 위한 개선이 아닌 고객이 원하는 핵심 요구사항을 개선하는 것, 고객 니즈를 정확히 이해하는 것이 중요
2. `데이터중심` : 감으로 판단하지 말고 데이터에 의해 객관적으로 판단하라
3. `프로세스중심` : 프로세스는 일하는 방식, 사고하는 방식, 부서 중심이 아닌 고객의 관점에서 최상의 프로세스를 구현하는 것이 문제해결방법론

데이터핸들링

- merge

```
pd.merge(left, right, how="inner", on=None, right_index=False, indicator=False)
# right_index: True라면, 오른쪽 DataFrame의 index를 merge Key로 사용
```

- join

```
pd.DataFrame.join(other, how="left", on=None
```