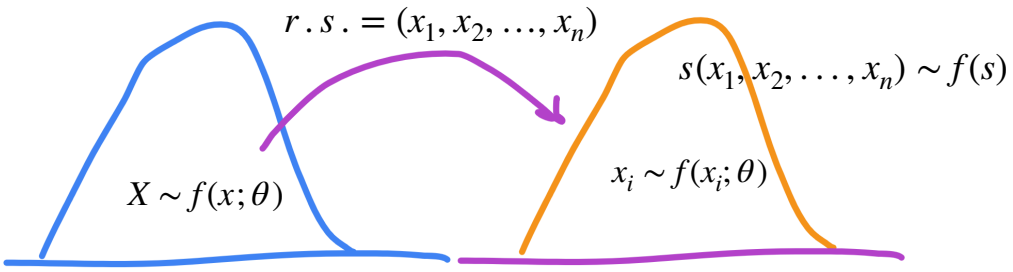
방법론 종류

비교방법

관심대상 개체 모집단의 관심특성(확률변수 X)의 데이터의 정보는 확률분포함수($f(x)$) 요약값 (중앙, 흩어짐 대표값)을 모수(θ)라 한다. 모집단으로부터 추출한 확률표본 데이터는 모집단의 모든 정보를 가진 축소 데이터이므로 확률표본의 표본 확률분포함수는 모집단 확률분포함수와 동일하다.



모집단에 대하여 알고 싶은 것은? (1) 확률분포함수 $f(x)$의 형태와 <=> 적합성 검정이라 한다. (2) 모수 θ이다. <=> 모수에 대한 추론(추정과 검정)이라 한다.

주요 관심모수

모집단의 요약값으로 가장 많이 사용되는 것은 중앙 위치, 흩어짐(산포)이다.

중앙위치

- 절대 중앙 : 산술평균 mean $\theta = \mu$ -> 만약 확률변수가 가질 수 있는 값이 (0,1) = binary(dichotomous)이면 산술평균은 비율이 된다. $\theta = p$: **[중심극한정리]**에 의해 통계량(표본평균, 표본비율)의 확률분포함수는 정규분포에 근사한다.
- 순서(상대) 중앙 : 중앙값 median $\theta = MD$ => 모수에 대한 추론은 통계량(표본 중앙값) 확률분포함수 (샘플링분포)를 알아야 한다. 그러나 순서 관련 통계량의 분포는 알 수 없으므로 "**비모수검정방법** non-parametric distribution free test"을 활용한다.

산포(흩어짐)

- 절대 흩어짐 : 분산 variance, 표준편차 standard deviation $\theta = \sigma^2$
- 상대(순서) 흩어짐 : 범위 range, 사분위 범위 mid-range $\theta = IQR$

독립인 2 집단 차이 비교 모수

평균차이 : $\theta = \mu_1 - \mu_2$ 비율차이 : $\theta = p_1 - p_2$ 분산차이 : $\theta = \dfrac{\sigma_1^2}{\sigma_2^2}$

통계량

확률표본 $(x_1, x_2, \ldots, x_n)$으로 계산된 값을 통계량 statistic $\{s(x_1, x_2, \ldots, x_n)$=확률표본의 함수$\}$ 이라 하고 모수의 점추정량(point estimator)으로 사용된다. 가장 좋은 추정량은 $MVUE$이다.

샘플링분포 sampling distirbution

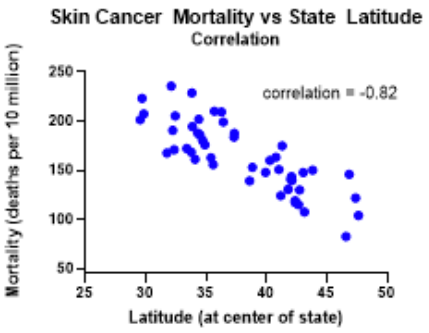
통계량의 확률분포함수(샘플링 분포)를 이용하여 모수에 대한 구간추정을 하거나 통계적 가설이라 한다.

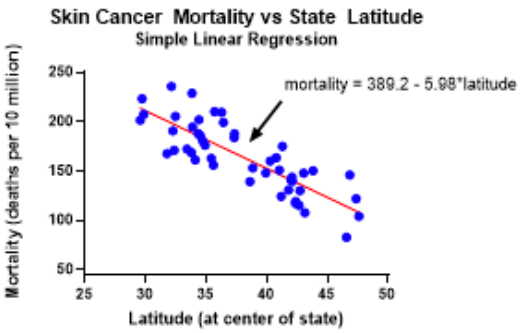
중심극한정리 : 모집단확률분포함수와 관계없이 표본크기가 충분히 크면 ($n > 20 \sim 30$) 표본평균, 표본비율, 표본합의 샘플링분포는 정규분포에 근사한다.

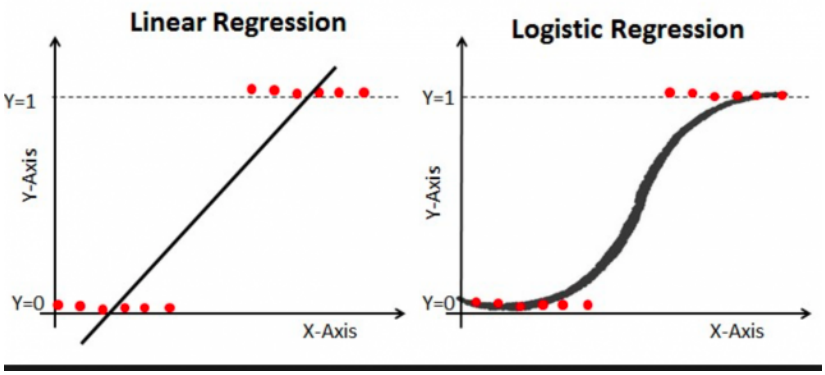
관계분석

(확률)변수들간의 상관(직선)관계 혹은 함수($y = f(x) + e, f$=선형)관계를 탐색한다.

종속변수(Y) 설명변수(X)	측정형	범주형
측정형	상관분석, 회귀분석	로지스틱 회귀분석
범주형	분산분석	로그-로그







개념

확증적 접근과 탐색적 접근

표본 데이터 수집이 완료되면 조사 목적에 적절한 통계 분석을 실시한다. 통계분석이라 함은 표본으로부터 계산된 통계량을 이용하여 ①모집단 모수(parameter: 모집단 평균, 비율, 분산)를 추정하거나(점 추정, 구간 추정) estimation ②모수에 대한 가설의 유의성을 검정하거나 hypothesis testing ③설정한 모형에 대한 유의성을 진단하게 된다.

빅데이터 시대에서는 통계량이나 시각화 분석을 통하여 1차 정보를 얻고 얻은 정보에 대한 확증적 분석을 실시하게 된다.

예제 데이터 : http://203.247.53.31/2015_Fall/D4BE/Bank2.csv

본 데이터는 "Managerial Statistics" - Keller 8th edition - 예제 데이터를 사용하였음

은행이 기업 대출 시 여성/남성 CEO에 따라 차별하는지 알아보기 위하여 조사한 자료이다. 작년 대출 신청한 남성 CEO 1,050명, 여성 CEO 115명에 대하여 승인여부, 대출 이자율, 비즈니스 타입(1=개인소유 proprietorship, 2=공동소유 patynership 3=기업형태 coporation), 기업 매출액(백만불), 기업 설립연수(년) 조사한 데이터이다.

```
import pandas as pd
bank=pd.read_csv('http://203.247.53.31/2015_Fall/D4BE/Bank2.csv')
bank.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1053 entries, 0 to 1052
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Gender    1053 non-null   object
 1   Rate      1053 non-null   float64
 2   Business  1053 non-null   int64
 3   Sales     1053 non-null   int64
 4   Age       1053 non-null   int64
```

일변량은 데이터의 변수를 하나씩 개별적으로 분석하는 방법이다.

- 수집된 데이터의 관심 특성, (확률)변수를 X라고 정의하자. (예) X=대출이자율, 매출액, 승인여부
- 모집단 전체 데이터이므로 $f(x;\theta)$는 알 수 있으나 유명 분포와 같은 수식 형태는 도출은 불가능하다.

데이터 종류

일변량분석의 경우 데이터의 종류에 따라 모수, 요약 통계량이 달라진다.

변수 종류	모수	시각 요약	숫자 요약
범주형(명목)	비율	바차트	비율, 최빈값
측정형 (구간, 순서, 비율)	평균, 분산	히스토그램 / 상자수염그림	평균, 중앙값, 분산, 사분위범위
시계열데이터	예측	시계열 도표	이동평균값, 지수평활값

- 집단이 2개인 경우도 일변량 분석에 포함한다. (예) 내야수와 외야수 연봉 차이
- 집단이 3개 이상인 경우에는 평균과 분산의 경우는 분산분석에서, 비율은 교차분석에서 다룬다.

분석 절차

1) 연구문제 정의

- 모집단 분석인지 표본 데이터 분석인지 결정한다. (필요 시 데이터 수집방법을 결정한다) 분석하려는 것이 대한 모수(관심 특성)을 결정한다.

2) 모집단 분석 (모집단 전수 데이터 확보 시)

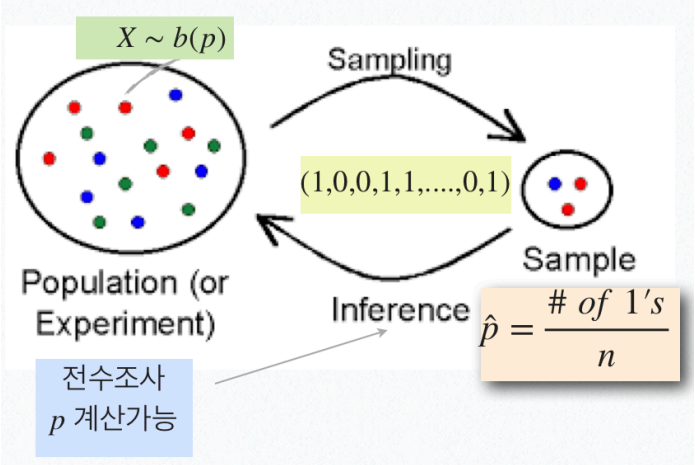
- 변수의 종류 : 측정형, 범주형인지 판단한다.
- 시각적 표현과 주요 통계량 계산으로 데이터를 요약한다.
- 필요 시 연구문제를 도출하여 3) 통계적 추론을 실시한다.

3) 통계적 추론

- 모집단으로부터 확률표본 데이터를 수집한다.
- 점 추정치(MVUE) 계산하고 필요 시 신뢰구간 계산한다
- 통계적 가설 설정 및 검정통계량 계산 및 유의확률 계산=>연구문제에 대한 통계적 판단을 결정한다.

모비율 추론 (일집단)

설정



- 확률변수 X는 성공, 실패 두 개의 결과만 갖는 베르누이 시행 결과
- 모집단 확률분포함수는 모수가 p인 베르누이분포임
- 확률표본의 i-번째 관측치 x_i의 값는 0(실패) 혹은 1(성공)이다.
- 모수 p는 모집단 전수조사 데이터의 경우에는 계산 가능하나 표본데이터에서는 MVUE 추정량으로 추정한다.
- 모비율에 대한 MVUE 추정량은 $\hat{p} = \frac{\#\ of\ success's}{n}$

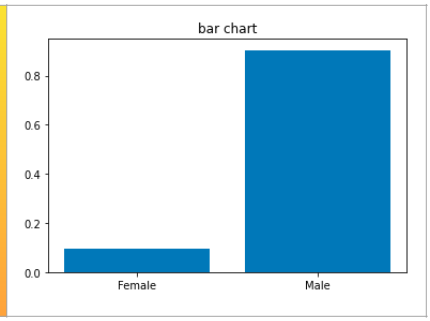
예제 데이터에서는 남자, 여자 승인률이 모수가 된다.

그래프 요약

```
x,f = np.unique(bank['Gender'],return_counts=True)
print (x, f/sum(f))
```

```
['Female' 'Male'] [0.09591643 0.90408357]
```

```
import matplotlib.pyplot as plt
plt.title('bar chart')
plt.bar(x,f/sum(f))
plt.show()
```



연구문제 정의

남자 승인율과 승인률 95%에 대한 신뢰구간을 구하시오.

작년 은행 기업 대출 승인율은 89%였다. 남자 승인율은 전체 승인율과 같은지 검정하시오.

모수 : $\theta = p$

점추정 for p

- 모비율의 MVUE 추정치는 표본 비율이다. $\hat{p} = \frac{\#\_of\_success}{\#\_of\_n}$
- 추정분산 $V(\hat{p}) = \frac{p(1-p)}{n}$
- 표본비율의 표준편차를 표준오차 (standard error)라 한다. $s(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

```
bank_freq=pd.crosstab(bank["Gender"],columns="count")
bank_freq
```

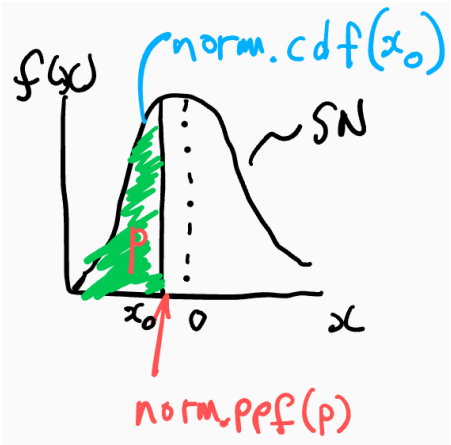
col_0	count
Gender	
Female	101
Male	952

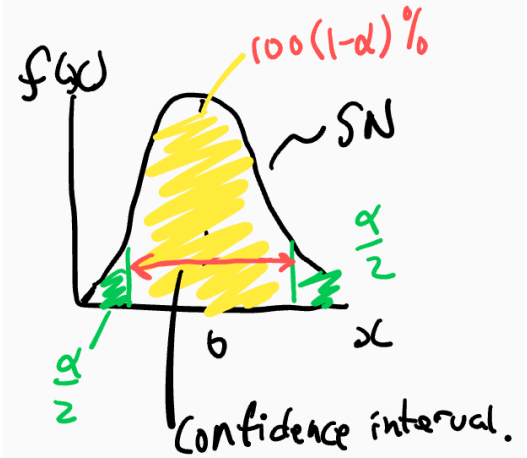
pd.crosstab() 빈도표를 만든다. 만약 columns= 에 범주형 변수를 쓰면 교차표가 나타난다.

```
nobs=1050
count=952
phat=count/nobs
phat
```

```
count    0.906667
```

남자 은행 대출 승인율은 90.7%이다.





구간추정 for p

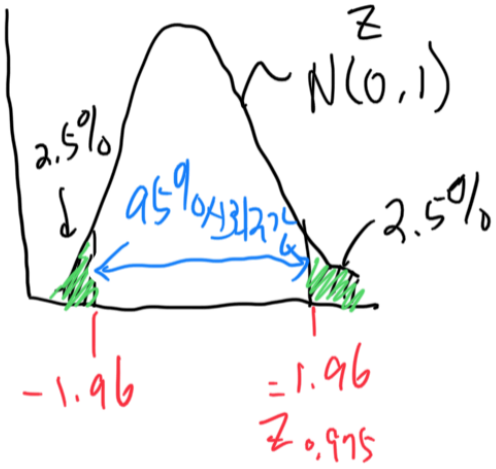
- 점추정치의 샘플링 분포 이용 : $\dfrac{\hat{p} - E(\hat{p}) = p}{s(\hat{p})} \sim z$,

  표준오차 (중심극한정리)

통계량의 표준편차를 표준오차 (standard error)라 한다. 구간 추정량을 계산할 때는 계산의 간편함을 위하여 표준오차의 p는 표본비율로 대체하여 사용한다.

$100(1-\alpha)\,\%$ 모비율 신뢰구간 :

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$



남자 은행 대출 승인율 95% 신뢰구간

```
import scipy.stats as st
lb=phat-st.norm.ppf(0.975,0,1)*((phat)*(1-phat)/n)**(0.5)
ub=phat+st.norm.ppf(0.975,0,1)*((phat)*(1-phat)/n)**(0.5)
lb,ub
```

```
(0.8890714080525564, 0.9242619252807769)
```

남자 은행 대출 승인율 95% 신뢰구간은 (88.9%, 92.4%)이다.

함수이용

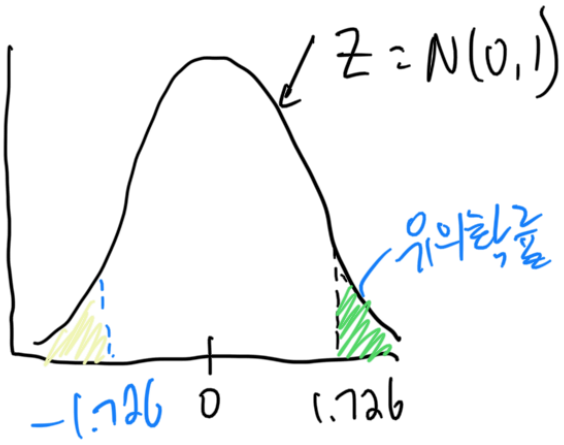
```
from statsmodels.stats.proportion import proportion_confint
proportion_confint(count, nobs, alpha=0.05, method='normal')
```

```
(0.8890714080525564, 0.9242619252807769)
```

검정

통계적 가설

- 귀무가설 : 남자 승인율은 은행 전체 대출 이자율 89%와 동일하가. $H_0 : p_0 = 0.89$
- 대립가설 : 남자 승인율은 89%와 같지 않다. $H_0 : p \neq 0.89$



검정통계량

$$TS = \frac{\hat{p} - p_0}{s(\hat{p})} \sim z, s(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$$

$$TS = \frac{\hat{p} - p_0}{\sqrt{(p_0)(1-p_0)/n}} = \frac{0.9067 - 0.89}{\sqrt{(0.89)(1-0.89)/1050}} = 1.726$$ (분모에 $\hat{p}$ 대신 p_0를 사용하는 이유는 귀무가설이 맞다는 가정 하에서 검정통계량을 구하기 때문임)

```
import numpy as np
p0=0.89
ts=(phat-p0)/np.sqrt(p0*(1-p0)/n)
ts
```

```
1.7260447582030933
```

유의확률계산

- $p\_value = 2 * (1 - Pr(z > |ts|)) = 0.084$ [양측대립가설]
- $p\_value = (1 - Pr(z > |ts|)) = 0.042$ [단측대립가설]
- 절대값을 취하는 이유는 항상 양의 값을

```
p_value=2*(1-st.norm.cdf(abs(ts),0,1))
print("검정통계량=%.2f | 유의확률=%.3f" %(ts, p_value))
```

```
검정통계량=1.73 | 유의확률=0.084
```

결론

유의확률이 0.084로 5%보다 커 귀무가설은 기각되지 않는다. 즉 남자 CEO의 은행 대출 승인율은 90.7%이나 전체 평균 대출 승인율 89%보다 높다고 할 수 없다.

만약 전체 대출 승인 이자율보다 높은가? 라고 연구문제를 정의하면 귀무가설은 동일한 대립가설이 $H_a : p > 0.89$ 단측으로 되어 유의확률은 0.042가 되어 귀무가설이 기각된다. 그러므로 양측대립가설의 경우 유의확률이 5%~10%인 경우에도 상승, 감소가 유의하다고 결론 내릴 수 있다

함수이용

결과가 다르다. 이유는? $TS = \dfrac{\hat{p} - p_0}{\sqrt{(p_0)(1-p_0)/n}}$ 계산 시 분모의 p_0 대신 $\hat{p}$ 를 대신 사용 -> 잘못된 계산 식을 사용하였음.

```
from statsmodels.stats.proportion import proportions_ztest
proportions_ztest(count,nobs,p0,alternative='two-sided')
```

```
(1.8565266430812573, 0.06337852903655086)
```

연습문제 https://vincentarelbundock.github.io/Rdatasets/csv/carData/TitanicSurvival.csv

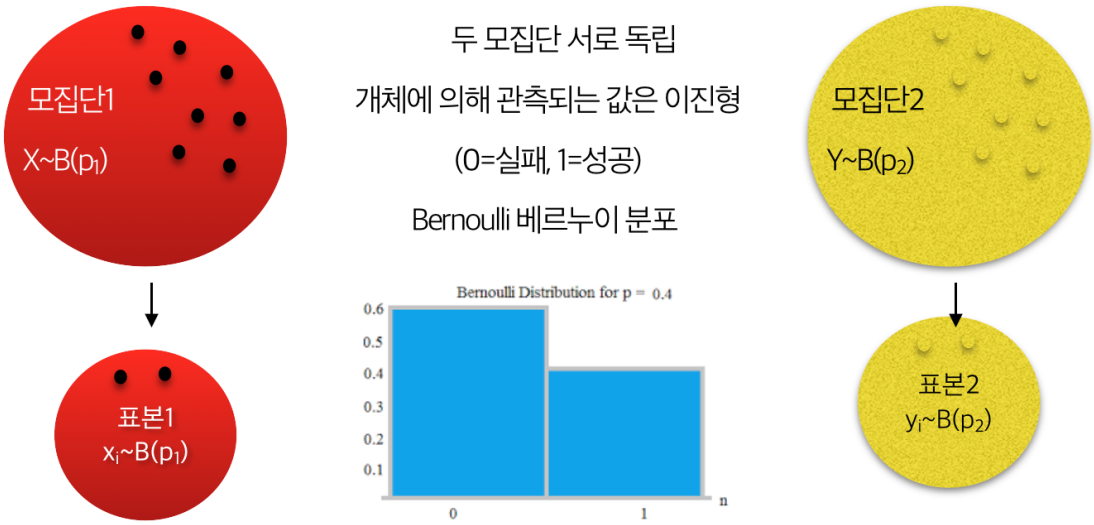
영국의 화이트 스타 라인이 운영한 북대서양 횡단 여객선이다. 1912년 4월 10일 영국의 사우샘프턴을 떠나 미국의 뉴욕으로 향하던 첫 항해 중에 4월 15일 빙산과 충돌하여 침몰하였다. 타이타닉 사고 탑승객 정보이다.

```
import pandas as pd
df=pd.read_csv('https://vincentarelbundock.github.io/Rdatasets/csv/carData/TitanicSurvival.csv')
df.info()
```

탑승객 생존율을 추정하고 95% 신뢰구간을 구하시오.

그 당시 배 사고 생존율이 40.5%였다면 타이타닉 사고가 더 많은 사망자를 내었다고 할 수 있는지 유의수준 5%에서 검정하시오.

독립인 두 모집단 비율 차이



확률표본 (iid) = 서로 독립이고 동일분포에서 추출

$$x_i \sim B(p_1) \qquad y_i \sim B(p_2)$$

$$x_1 = 0, x_1 = 1, x_3 = 1, \ldots, x_n = 0 \quad \text{데이터} \quad y_1 = 1, y_2 = 1, y_3 = 0, \ldots, y_m = 0$$

연구문제 정의

은행이 남녀 CEO 대출 승인율 측면에서 차별하는지 알아보기 위하여 남녀 승인율 차이를 검정해보자.

여성 CEO 대출 승인율 = p_1, 남성 CEO 대출 승인율 = p_2

모수 : $\theta = p_1 - p_2$

점추정 $\hat{\theta} = \hat{p}_1 - \hat{p}_2$

- 추정오차 $V(\hat{p}_1 - \hat{p}_2) = \dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$
- 표준오차 : 추정분산의 양의 제곱근 $s(\hat{p}_1 - \hat{p}_2) = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$

- 1집단(여성 CEO) 표본크기 n_1, 2집단(남성 CEO) 표본크기 n_2
- 1 집단(여성 CEO 대출 신청 중) 성공(대출승인) 회수 x, 2 집단(남성 CEO 대출 신청 중) 성공(대출승인) 회수 y,
- 점추정 : $\hat{\theta} = \hat{p}_1 - \hat{p}_2$=0.878-0.907=-0.0284
- 여성 CEO 대출 승인율 추정값 : $\hat{p}_1 = \dfrac{no.\ of\ 1s}{n_1} = \dfrac{x}{n_1}$=$102/115 = 0.878$

• 남성 CEO 대출 승인율 추정값 : $\hat{p}_2 = \frac{no.\ of\ 1s}{n_2} = \frac{y}{n_2}$=952/1050 = 0.907

```
import numpy as np
count=np.array([101,952])
nobs=np.array([115,1050])
phat=count/nobs
phat
```

```
array([0.87826087, 0.90666667])
```

구간추정

샘플링 분포 : $\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s(\hat{p}_1 - \hat{p}_2)} \sim z$ **[중심극한정리]**

표본비율 차이($\hat{\theta} = \hat{p}_1 - \hat{p}_2$) 표준오차(SE) : $s(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

추정치 $\hat{p}_1 - \hat{p}_2$의 표준오차 $s(\hat{p}_1 - \hat{p}_2)$ 계산 시 계산 간편을 위하여 모수 값 대신 추정값을 사용하여 신뢰구간을 구한다.

$100(1-\alpha)\%$ 모비율 차이($p_1 - p_2$) 신뢰구간 :

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

여성 CEO 대출승인율 = 87.8%, 남성 CEO 대출승인율 90.7% 차이 95% 신뢰구간은 (-0.091, 0.034)으로 0을 포함하고 있으므로 승인율의 차이는 없다.

```
se=np.sqrt(phat[0]*(1-phat[0])/nobs[0]+phat[1]*(1-phat[1])/
nobs[1])
import scipy.stats as st
lb=(phat[0]-phat[1])-st.norm.ppf(0.975,0,1)*se
ub=(phat[0]-phat[1])+st.norm.ppf(0.975,0,1)*se
lb,ub
```

```
(-0.0907043104888855, 0.03389271628598711)
```

함수 이용

남여 각각의 모비율 신뢰구간을 구해 준다.

```
from statsmodels.stats.proportion import proportion_confint
proportion_confint(count, nobs, alpha=0.05, method='normal')
```

```
(array([0.81849874, 0.88907141]), array([0.938023  , 0.92426193]))
```

통계적 가설검정

- 귀무가설 : 여성 CEO 대출 승인율과 남성 CEO 대출 승인율은 동일하다. $H_0 : p_1 = p_2$
- 대립가설 : 은행은 대출 승인율 측면에서 여성 CEO를 차별한다. $H_a : p_1 < p_2$
- 검정통계량 : $TS = \frac{\hat{p}_1 - \hat{p}_2 - 0}{s(\hat{p}_1 - \hat{p}_2)} \sim z,\ s(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_0(1-p_0)}{n_1} + \frac{p_0(1-p_0)}{n_2}}$
- p_0 통합 비율(pooled proportion) : $p_0 = \frac{x+y}{n_1+n_2}$

추정치 $\hat{p}_1 - \hat{p}_2$의 표준오차 $s(\hat{p}_1 - \hat{p}_2)$ 계산 시 가설검정에는 귀무가설의 설정값, 즉 두 모수 값이 동일하다는 가정에 의해 추정되는 값을 넣는다.

```
phat0=sum(count)/sum(nobs)
phat0 #통합비율
```

```
0.903862660944206
```

```
se2=np.sqrt(phat0*(1-phat0)/nobs[0]+phat0*(1-phat0)/nobs[1])
ts=(phat[0]-phat[1])/se2
p_value=1-st.norm.cdf(abs(ts),0,1)
print("검정통계량=%.2f | 유의확률=%.3f" %(ts, p_value))
```

```
검정통계량=-0.98 | 유의확률=0.163
```

결론

유의확률 0.163로 유의수준 5%보다 크므로 귀무가설은 채택되어 은행은 대출 승인율 측면에서 남성, 여성 CEO 차별하지 않는다.

함수이용

```
from statsmodels.stats.proportion import proportions_ztest
proportions_ztest(count,nobs,alternative='smaller')
```

```
(-0.9810481340485643, 0.1632845027581592)
```

연습문제 https://vincentarelbundock.github.io/Rdatasets/csv/carData/TitanicSurvival.csv

타이타닉 사고 시 여성은 남성에 비해 생존 가능성 우대를 받았는지 유의수준 5%에서 검정하시오.

세미소사와 맥과이어 홈런 경쟁으로 인하여 여성 팬이 증가하였다고 주장한다. 이를 알아보기 위하여 CNN/ USA 이 다음 조사를 하였다. 1995년 1008명 여성 중 413이 팬이라고 대답했고, 홈런 경쟁이 있은 1998년에는 1082 여성 중 681명이 팬이라고 답하였다. 이를 이용하여 주장에 대해 답하시오.

짝진 집단 비율 차이 McNemar 검정

연구문제(시나리오)

- 동일한 개체로부터 이진형(성공, 실패) 변수를 서로 다른 기간(before - after)에 측정하여 프로그 램 효과가 있는지 알아보는 방법이다.
- Bland (2000) 1319명 어린이에 대하여 12살에 독감에 걸릴 가능성은 나이가 14살이 되면 높아지는지, 낮아지는지 알아보기 위하여 조사한 결과이다. 즉 1,319명 어린이는 동일하며 12살에 조사하고 14살에 다시 한 번 조사한 결과이다. https://www.medcalc.org/manual/ mcnemartest2.php

Severe colds at age 12	Severe colds at age 14		Total
	Yes	No	
Yes	212 A	144 B	356
No	256 C	707 D	963
Total	468	851	1319

각 결과에 대한 주별확률 밀도함수는 동일하다 <=> 감염율의 변화는 없다.

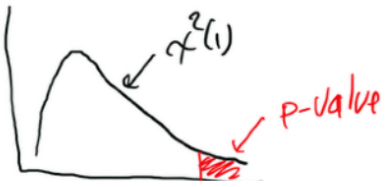
P(A)+P(B)(12 살에 걸린 독감 걸린 사람 비율)=P(A)+P(C)(14 살에 걸린 독감 걸린 사람 비율)
P(C)+P(D)(12 살에 걸린 독감 **안** 걸린 사람 비율)=P(B)+P(D)(14 살에 걸린 독감 **안** 걸린 사람 비율)
그러므로 귀무가설은 $P(B) = P(C)$

통계적 가설

- 귀무가설 : 나이가 올라가도 독감 걸릴 가능성은 동일하다. $P(B) = P(C)$
- 대립가설 : 나이가 올라가도 독감 걸릴 가능성은 달라진다.

점추정치 및 샘플링 분포

$$TS = \frac{(B-C)^2}{B+C} = \frac{(144-256)^2}{144+256} = 31.36 \sim \chi^2(1)$$



=CHISQ.DIST.RT(31.36,1)

D	E
0.00000002	

- 유의확률 : <0.001, 귀무가설 기각

결론

- 12살 독감 걸렸던 대상자가 14살에 걸릴 확률은 40%(=212/356)로 12살에 안 걸렸던 대상자가 14살 독감 걸릴 확률은 26.6%(=256/963)보다 유의적으로 높다.
- 결론적으로 독감에 걸렸던 대상자가 나이가 들어도 다시 걸릴 가능성이 높아진다.

모비율 소표본 $min(np, n(1-p)) \leq 5$

분포이용

모비율 검정 시 검정통계량의 분포는 정규분포를 가정한다.

이는 중심극한 정리에 의한 것으로 [중심극한정리 central limit theorem]모집단의 분포가 어떠하든지 표본평균, 표본합 (표본비율도 표본평균과 동일 개념) 분포는 정규분포를 따른다.

소표본인 경우는 이항분포를 이용하여 가설을 검정한다. 모집단의 개체가 성공(성공 확률이 p), 실패의 결과만 있으므로 확률표본은 베르누이 시행과 동일하다. 그러므로 표본크기 n 확률표본으로부터 구한 성공 개체의 수 합은 이항분포 $B(n,p)$를 따른다.

학생 흡연 비율이 20% 미만이라고 발표했다. 맞 는지 알아보기 위하여 학생 20명을 확률 추출하여 흡연여부를 알아본 결과 3명이 흡연하고 있다고 조사 되었다. 발표가 맞는지 검정하시오.

귀무가설 : $H_0 : p = 0.2$

대립가설 : $H_0 : p < 0.2$

$min(20 * 0.2, 20 * 0.8) = 4 < 5$이므로 중심극한정리 적용이 불가능하다. 대신 흡연자 수는 이항분포 (20, p=0.2) [귀무가설이 맞다는 가정 하에 구하게 된다] 따르므로 유의확률은 다음과 같다.

유의확률 : $P(\sum X_i \leq 3 \,|\, sum X_i \sim B(n = 20, p = 0.2))$ =0.42

귀무가설 기각할 수 없음

```
import scipy.stats as stat
stat.binom.cdf(3,20,0.2)
```

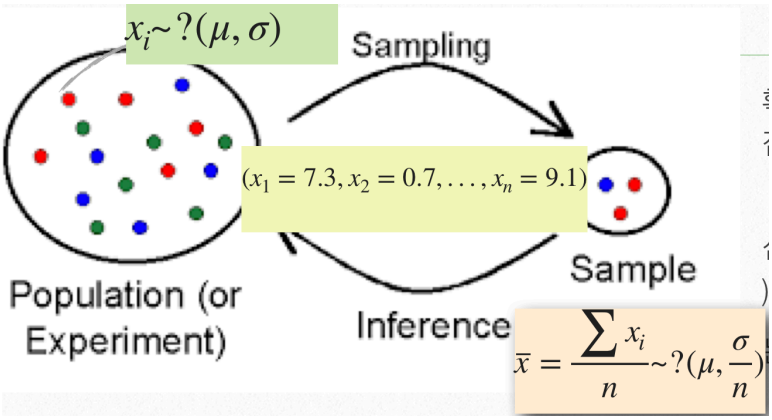
```
0.41144886195656954
```

Wilson 통계량

표본크기 n에 비해 성공회수 $\sum x_i = x$ 가 매우 작은 경우 비율 추정치는 $\hat{p} = \frac{x}{n}$ 대신 $\hat{p} = \frac{x+2}{n+4}$를 사용하고 추정방법은 모두 동일하다.

모평균 one population mean 추론

설정



- 확률변수 X는 측정형이며 모집단 확률분포함수 $f(x;\theta)$이다.
- 그러므로 확률표본 $(x_1, x_2, \ldots, x_n)\ iid\ \sim f(x;\theta)$이다. 서로 독립 independently 이고 indentical 동일 모집단 분포를 따른다.
- 모집단 확률분포함수 관심모수는 $\theta = \mu$(모평균)이고 모르는 모집단 분산 σ^2은 부가 ancially 모수이다.. 부가모수는 모르는 값이지만 추정 관심이 없다.
- 모수(μ, σ)은 모집단 전수조사 데이터의 경우에는 계산 가능하나 표본데이터에서는 MVUE 추정량으로 추정한다.
- MVUE 추정량 : 표본평균 $\hat{\mu} = \bar{x} = \frac{\sum x_i}{n}$ / 표본분산 $\hat{\sigma^2} = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ 이다.

연구문제

작년 기업 CEO 대출 이자율이 1.33%였다. 올해 대출 이자율이 낮아졌는지 알아보기 위하여 가설검정을 하고 올해 대출 평균 이자율 95% 신뢰구간을 구하시오.

그래표 요약 [참고 : http://wolfpack.hnu.ac.kr/Fall_2020/lecturenote/sm_graphic.pdf]

[참고 : http://wolfpack.hnu.ac.kr/Fall_2020/lecturenote/sm_good_of_fits.pdf]

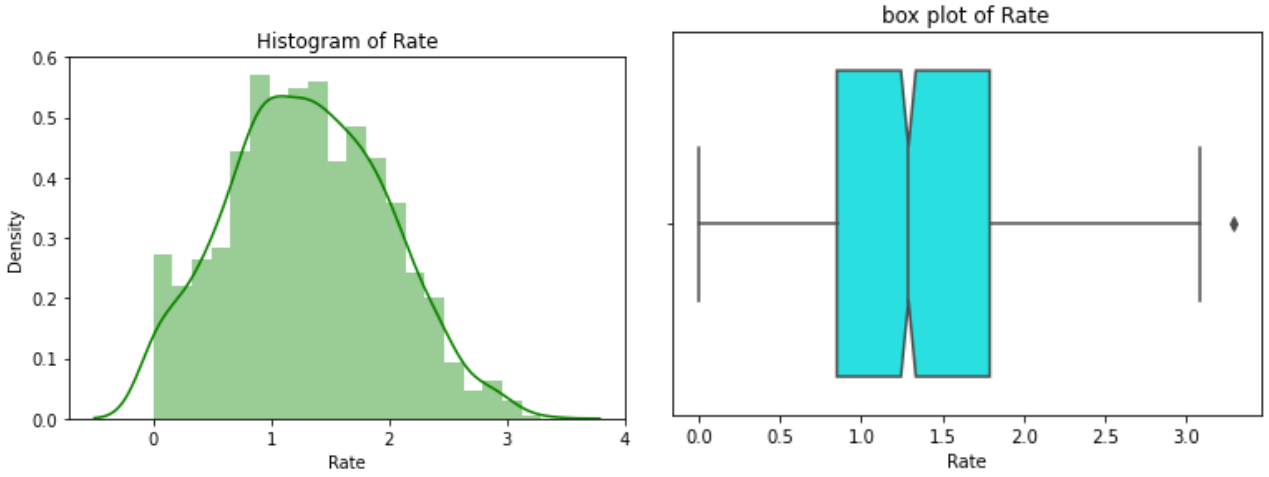
[참고 : http://wolfpack.hnu.ac.kr/Fall_2020/lecturenote/sm_statistic_sampling_dist.pdf]

히스토그램

- 확률분포함수 형태 - 모집단 확률분포함수와 동일 (좌로 치우침, 우로 치우침, 죄우대칭)
- 봉우리 (최빈값)의 개수를 알 수 있다. : 두 개의 서로 다른 집단 데이터의 히스토그램을 그리는 경우 봉우리가 2개 나타나는 경우가 발생 - 상자 수염 그림으로는 판별 불가

```
import seaborn as sns
import matplotlib.pyplot as plt
ax=sns.distplot(bank["Rate"],label="interest
rate",bins=20,color='green')
ax.set_title('Histogram of Rate')
plt.show()
```

좌우대칭에 가까우나 약간 우로 치우친 형태



상자 그림 box-whisker plot

- 확률분포함수 형태 - 모집단 확률분포함수와 동일 (좌로 치우침, 우로 치우침, 죄우대칭) 단 봉우리는 모른
- 이상치 outliers : 두개의 가상선 $Q_1 - 1.5IQR$, $Q_3 + 1.5IQR$을 벗어나는 관측값
- 극심 이상치 : 1.5대신 3을 사용

```
import seaborn as sns
import matplotlib.pyplot as plt
ax=sns.boxplot(x="Rate",notch=True,data=bank,color='cyan')
ax.set_title('box plot of Rate')
plt.show()
```

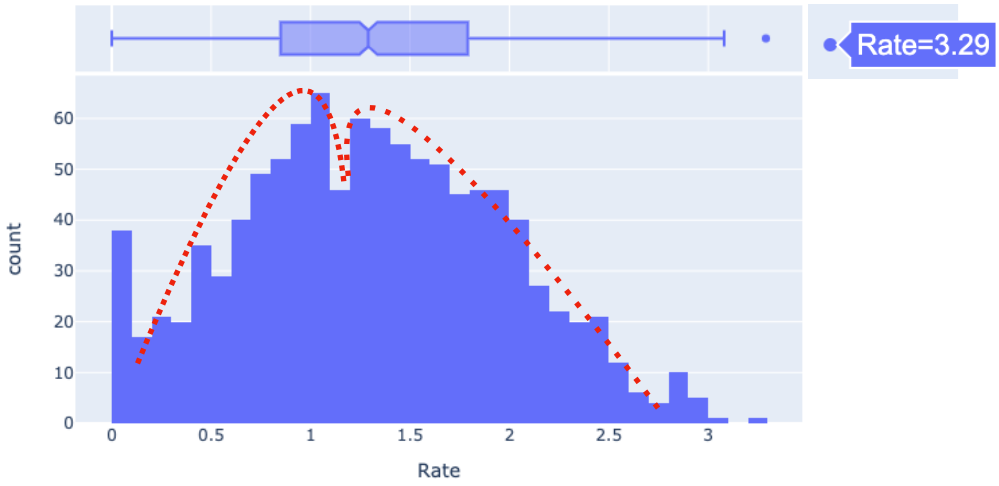
이상치 한 개가 존재한다.박스 안 상자의 크기가 우측이 살짝 크고 왼쪽 수염보다 오른쪽 수염 길이가 길므로 우로 약간 치우친 형태이다.

상자그림_히스토그램 한 번에

```
import plotly.express as px
fig=px.histogram(bank,x='Rate',marginal="box")
fig.show()
```

우로 치우침을 보이고 한 개의 이상치가 있는 것으로 판단됨 - 이자율 3.29는 이상치로 판단되어 향후 분석에서는 제외함

남여 대출 이자율이 합쳐 있어 쌍봉(봉우리 2개) 형태를 갖는다. 만약 남여 집단이 혼재되어 있는 줄 모르는 상태에서 이런 현상이 생기면 집단 변수를 찾아야 하는 어려움이 발생한다.



추정

모평균 μ 점 추정

MVUE 표본평균 : $\bar{X} = \dfrac{\sum x_i}{n}$

- 추정분산 : $V(\bar{X}) = \dfrac{\sigma^2}{n}$ => 표준오차 : $s(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$, 만약 σ 모르면 표준편차 s 로 대체한다.

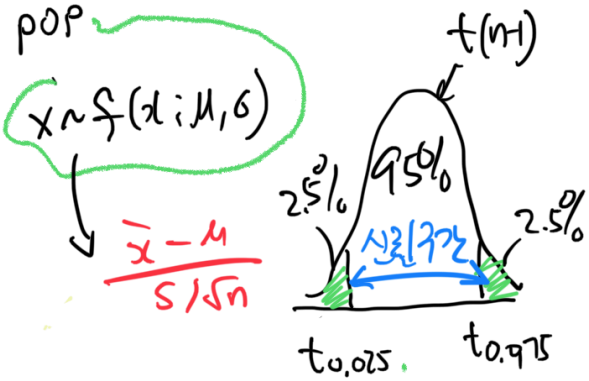
```
import numpy as np
n=bank.loc[bank['Rate']<3.29,'Rate'].count()
xbar=bank.loc[bank['Rate']<3.29,'Rate'].mean()
sd=bank.loc[bank['Rate']<3.29,'Rate'].std()
se=bank.loc[bank['Rate']<3.29,'Rate'].std()/np.sqrt(n)
print('smaple size=%d, sample mean=%.3f sample std=%.4f'%(n,xbar,sd))
```

```
smaple size=1052, sample mean=1.301 sample std=0.6665
```

구간 추정

[샘플링 분포]

- 대표본($n \geq 20 \sim 30$) : 표본평균의 샘플링분포 : [중심극한정리] $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim N(0,1) = z$
- 소표본 : 모집단 정규분포 가정 하에 $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(df = n-1)$



[$100(1-\alpha)$ % 구간추정]

$$(\bar{x} - t_{1-\alpha/2;n-1}\frac{s}{\sqrt{n}},\ \bar{x} + t_{1-\alpha/2;n-1}\frac{s}{\sqrt{n}})$$

- 대부분의 통계 소프트웨어는 표본 크기에 관계없이 t-분포를 사용한다.

t-분포 : 표준정규분포와 동일한 좌우 대칭 형태이며 평균도 0으로 동일하다. 자유도가 n인 경우 t-분포의 분산은 $\frac{n}{n-2}$으로 표준정규분포 1보다 크므로 꼬리부분이 표준정규분포보다 두텁다. 자유도가 커지면 분산이 1로 근사하여 t-분포는 표준정규분포에 근사한다.

자유도(degree of freedom) : 데이터가 가진 정보, n개 데이터는 서로 독립적으로 정보를 가지고 있어 자유도는 n이나 평균을 계산한 후에는 하나의 데이터를 잃어도 평균으로부터 그 데이터 값을 추정할 수 있다. 그러므로 모집단 평균의 분포의 자유도는 (n-1)이 된다.

```
import scipy.stats as stat
lb=xbar-stat.t.ppf(0.975,n-1)*se
ub=xbar+stat.t.ppf(0.975,n-1)*se
print('95%% 모평균 신뢰구간=(%.f, %.3f)' %(lb,ub))
```

```
95% 모평균 신뢰구간=(1.261, 1.342)
```

함수 이용

```
import scipy.stats as stat
stat.t.interval(0.95,n-1,xbar,se)
```

```
(1.2611488660688683, 1.34179790199197)
```

통계적 가설검정

- 귀무가설 : 올해 은행 대출 이자율은 작년 1.33%와 동일하다. $H_0 : \mu = \mu_0 = 1.33$
- 대립가설 : 올해 은행 대출 이자율은 작년 1.33%보다 낮아졌다.. $H_a : \mu < 1.33$
- 검정통계량 : $TS = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(df = n-1)$

```
import numpy as np
import scipy.stats as stat
mu0=1.33
ts=(xbar-mu0)/se
pvalue=1-stat.t.cdf(np.abs(ts),n-1)
print('검정통계량=%.2f, 유의확률=%.3f)' %(ts,pvalue))
```

```
검정통계량=-1.39, 유의확률=0.083)
```

함수이용

```
import scipy.stats as stat
ts,pvalue=stat.ttest_1samp(bank.loc[bank['Rate']<3.29,'Rate'],1.33)
print('검정통계량=%.2f, 유의확률=%.3f)' %(ts,pvalue/2))
```

```
검정통계량=-1.39, 유의확률=0.083)
```

결론

유의확률 0.083로 유의수준 5%보다 크므로 귀무가설은 채택되어 올해 은행 대출 이자율은 1.301%로 작년(1.33%)에 비해 낮아졌으나 유의적으로 낮아지지 않았다.

연습문제 http://wolfpack.hnu.ac.kr/Stat_Notes/elem_stat/Stat_methods/newspaper.csv

도시 환경과에서 신문 수거 업체를 활용하고자 한다. 수거 업체는 가구당 하루 2파운드 이상 배출해야 수 거하여 수익이 발생한다고 하였다. 환경과에서 수거 업체의 수익을 보장할 수 있는지 알아보기 위하여 148가구에 대하여 조사한 자료이다.

신문수거업체를 활용하는 것이 적절한지 1)상자그림과 히스토그램을 그리고(이상치가 있는 경우 제외) 2)95% 신뢰구간을 구하고 3)유의수준 5%에서 가설 검정하시오.

연습문제 : Major League Baseball Data from the 1986 and 1987 seasons.

https://vincentarelbundock.github.io/Rdatasets/csv/ISLR/Hitters.csv

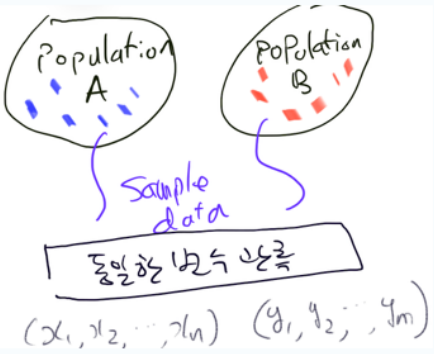
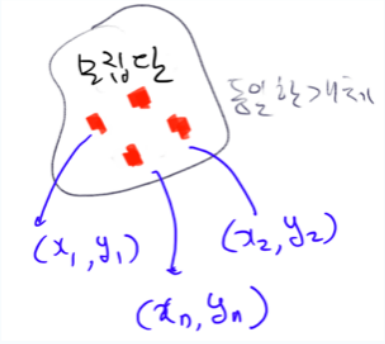
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 322 entries, 0 to 321
Data columns (total 21 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  322 non-null    object
 1   AtBat       322 non-null    int64
 2   Hits        322 non-null    int64
 3   HmRun       322 non-null    int64
 4   Runs        322 non-null    int64
 5   RBI         322 non-null    int64
 6   Walks       322 non-null    int64
 7   Years       322 non-null    int64
 8   CAtBat      322 non-null    int64
 9   CHits       322 non-null    int64
 10  CHmRun      322 non-null    int64
 11  CRuns       322 non-null    int64
 12  CRBI        322 non-null    int64
 13  CWalks      322 non-null    int64
 14  League      322 non-null    object
 15  Division    322 non-null    object
 16  PutOuts     322 non-null    int64
 17  Assists     322 non-null    int64
 18  Errors      322 non-null    int64
 19  Salary      263 non-null    float64
 20  NewLeague   322 non-null    object
dtypes: float64(1), int64(16), object(4)
```

1. 1987년 연봉 높은 순으로 20명을 선택하여 연봉 점도표를 그리시오.
2. 타석(AtBat) 상위 25% 이상인 선수들만 선택하시오. 데이터프레임 baseball 에 저장하시오.
3. 타율(batting average)을 계산하시오. BA=Hits/AtBat
4. 타율에 대한 히스토그램과 상자그림을 동시에 그리고 치우침 등 분포의 형태와 이상치 진단하시오. 이상치가 존재하면 제외하고 향후 분석을 시작하시오.
5. 타율의 95% 신뢰구간을 구하시오.
6. 선수들의 타율이 2할 5푼 이상인지 유의수준 5%에서 검정하시오.

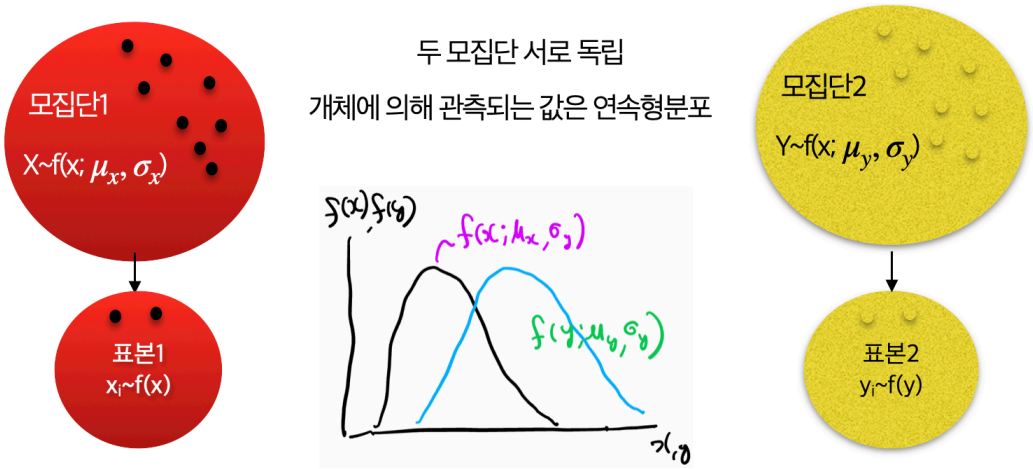
독립인 모평균 차이 추론 two independent population mean

개념

비교하는 2개의 집단은 "독립" independent, "짝진" paired 으로 나뉜다. 독립과 짝진의 구별은 관측 대상이 동일 개체인지 상이한 개체인지에 의해 결정된다.

독립집단 경우	짝진집단 경우
	
관측변수 X=Y 동일함	
서로 다른 개체로 부터 관측한다.	동일개체로부터 관측된다.
관측치 개수 n, m 크기는 같을 필요 없음	관측치 개수는 n개 항상 동일하다.
경영학 전공자와 통계학 전공자의 일주일 공부 시간 차이를 알고자 한다. 경영학 전공자 30명, 통계학 전공자 20명을 임의 추출(확률표본)하여 공부시간을 측정하였다.	경영학 전공자와 통계학 전공자의 일주일 공부 시간 차이를 알고자 한다. 그런데. 학점 군에 따른 공부시간 차이를 고려하여 (4.5~4.3), (4.3~4.1), (4.1~3.9), ... 각 구간에서 한 명씩 임의 추출하였다.

독립집단



$x_1 = 11.2, x_2 = 15.3, \ldots, x_n = 11.1$ 데이터 $y_1 = 12.1, y_2 = 12.3, \ldots, y_m = 11.11$

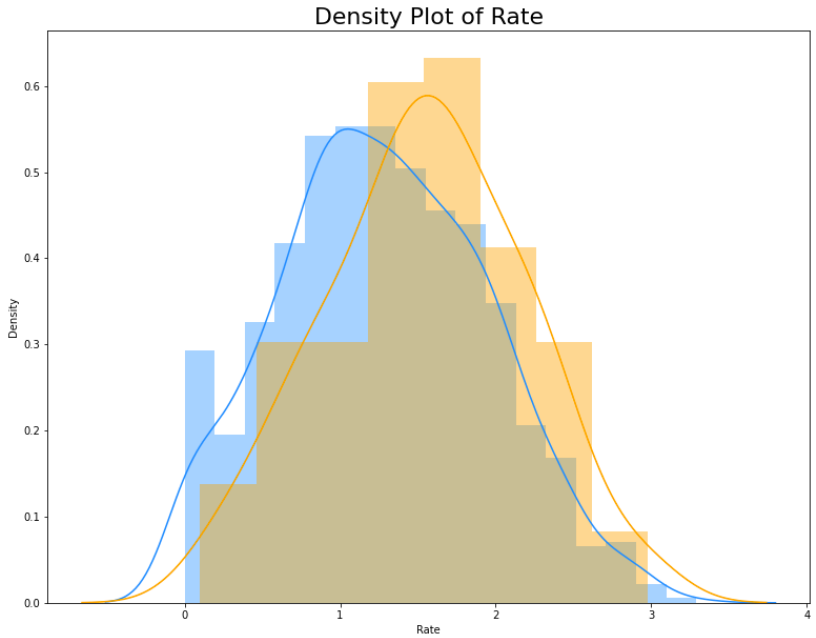
확률표본 (iid) = 서로 독립이고 동일분포에서 추출

연구문제

은행이 대출 이자율 측면에서 성 차별한다고 여성 CEO들이 주장하였다. 이 주장이 맞는지 유의수준 5%에서 검정하시오. 은행이 차별한다는 것은 여성 CEO 평균이자율이 남성 CEO 이자율보다 높은가에 대한 검정

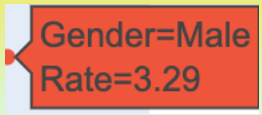
시각적 그래프

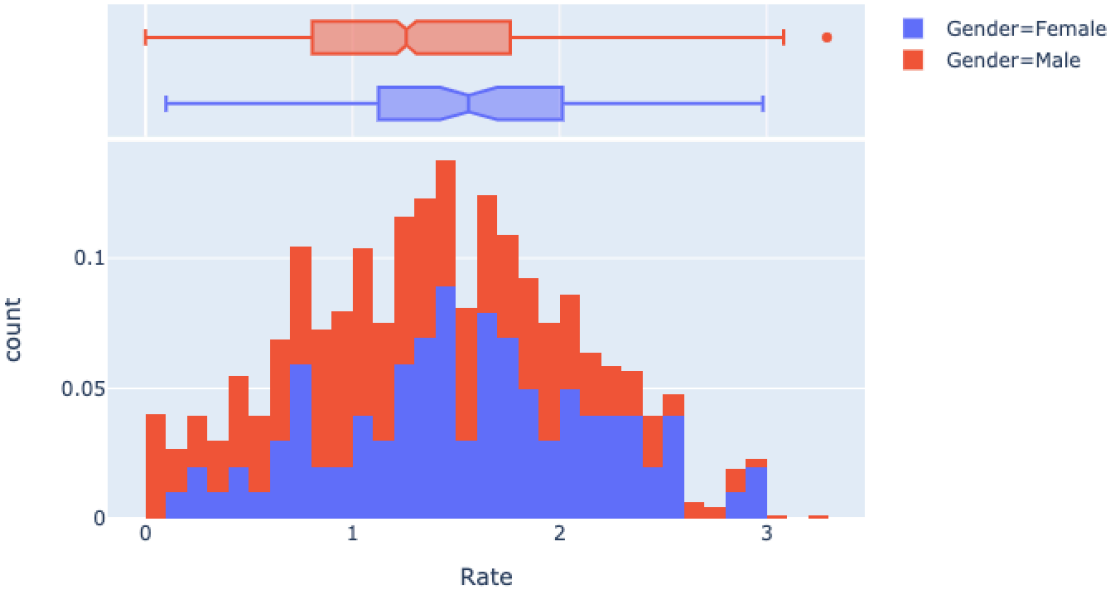
```
import matplotlib.pyplot as plt
import seaborn as sns
df=bank
plt.figure(figsize=(13,10))
sns.distplot(df.loc[df['Gender']=='Male',"Rate"],color="dodgerblue", label="american")
sns.distplot(df.loc[df['Gender']=='Female',"Rate"],color="orange", label="japanese")
plt.title('Density Plot of Rate',fontsize=22)
plt.show()
```



```
import plotly.express as px
fig = px.histogram(bank,x='Rate',color='Gender',marginal="box",histnorm='probability')
fig.show()
```

두 남녀 이자율 분포는 유사하며(분포의 형태는 당연히 동일한) 남자 CEO의 경우 이상치가 한 개 존재함





점추정 모평균 차이 $\theta = \mu_x - \mu_y$

- 집단 1 모평균 μ_x 집단 2 모평균 μ_y
- 표본크기 : n, m

점추정 $\hat{\theta} = \bar{x} - \bar{y}$ MVUE

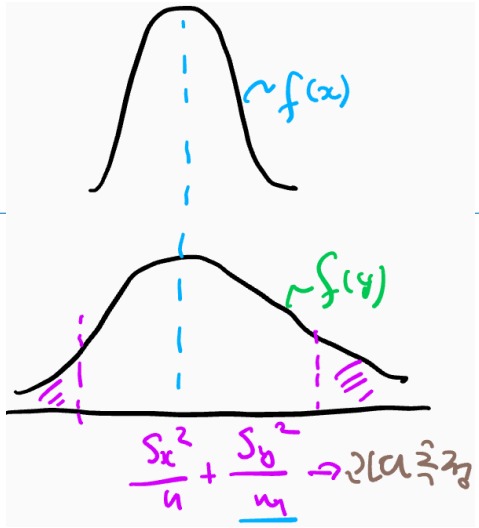
$$\hat{\mu}_x = \bar{x} = \frac{\sum x_i}{n}, \hat{\mu}_y = \bar{y} = \frac{\sum y_i}{m}, E(\hat{\theta}) = E(\bar{x} - \bar{y}) = \mu_x - \mu_y$$ 불편추정량

추정분산 $V(\bar{x} - \bar{y})$

. $V(\bar{x} - \bar{y}) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$ => 표준오차 : $s(\bar{x} - \bar{y}) = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$

$\hat{\theta} = \bar{x} - \bar{y}$ 샘플링분포

- 중심극한 정리 : $\bar{x} \sim N(\mu_x, \frac{\sigma_x}{\sqrt{n}}), \bar{y} \sim N(\mu_y, \frac{\sigma_y}{\sqrt{m}})$
- 정규분포 가법성에 의하여 $\bar{x} - \bar{y} \sim N(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}})$
- 통합분산 : $s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$



모평균 검정 시 모분산 동일한지 검정해야 하는 이유

두 독립인 모집단의 평균이 동일하더라도 분산이 서로 다르면 왼쪽 분산이 큰 데이터 확률 추출 시 안쪽 부분이나 오른쪽 부분에서 좀 더 많은 개수의 표본이 추출된다면 평균이 왜곡된다.

그리고 검정통계량으로 사용되는 분산이 과다 추정될 가능성이 있어 이를 고려해야 한다.

기초통계량 구하기

```
df=bank[bank['Rate']<3.29] #이상치 1개 제외
df.groupby(by='Gender').describe()['Rate']
```

Gender	count	mean	std	min	25%	50%	75%	max
Female	101.0	1.545446	0.636721	0.1	1.130	1.56	2.01	2.98
Male	951.0	1.275563	0.664710	0.0	0.805	1.26	1.76	3.08

등분산 가정 만족 시 샘플링 분포

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{s_p^2(\frac{1}{n} + \frac{1}{m})}} \sim t(N+m-2), s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

등분산 가정 무너지면 샘플링 분포

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}} \sim t(Welch's\ df), \text{자유도 복잡}$$

$$\frac{(\frac{s_x^2}{n-1}+\frac{s_y^2}{m-1})^2}{\frac{s_x^{2^2}}{n^2(n-1)}+\frac{s_y^{2^2}}{m^2(m-1)}}$$

모평균 차이 $100(1-\alpha)\,\%$ 신뢰구간

$$(\bar{x}-\bar{y}) \pm t(\alpha/2, n+m-2)\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}$$

```
import numpy as np
mean_x=df.loc[df['Gender']=='Female','Rate'].mean()
mean_y=df.loc[df['Gender']=='Male','Rate'].mean()
std_x=df.loc[df['Gender']=='Female','Rate'].std()
std_y=df.loc[df['Gender']=='Male','Rate'].std()
n=df.loc[df['Gender']=='Female','Rate'].count()
m=df.loc[df['Gender']=='Male','Rate'].count()
se_x=std_x/np.sqrt(n)
se_y=std_y/np.sqrt(m)
```

```
lb=(mean_x-mean_y)-stat.t.ppf(0.975,n+m-2)*np.sqrt(std_x**2/n+std_y**2/m)
ub=(mean_x-mean_y)+stat.t.ppf(0.975,n+m-2)*np.sqrt(std_x**2/n+std_y**2/m)
print('95%% 모평균 차이 신뢰구간=(%.4f, %.4f)' %(lb,ub))
```

```
95% 모평균 차이 신뢰구간=(0.1386, 0.4012)
```

함수 이용

두 모평균 차이에 대한 신뢰구간 함수는 존재하지 않는다. 활용성이 매우 낮기 때문이다. 대신 각각의 신뢰구간을 구하여 재시하는 것이 적절하다.

```
import scipy.stats as stat
stat.t.interval(0.95,n-1,mean_x,se_x),
stat.t.interval(0.95,m-1,mean_y,se_y)
```

```
((1.419748922072141, 1.6711421670367712),
 (1.2332622576860879, 1.3178628737545037))
```

등분산 검정 $H_0 : \sigma_x^2 = \sigma_y^2$

검정통계량 : $TS = \frac{max(s_x^2, s_y^2)}{min(s_x^2, s_y^2)} \sim F(df1 = df_{num}, df2 = df_{den})$

유의확률이 41.3%로 유의수준 5%보다 크므로 귀무가설이 채택되어 등분산 가정이 만족한다.

```
import scipy.stats as st
ts,pvalue=st.levene(df.loc[df['Gender']=='Male','Rate'],df.loc[df
['Gender']=='Female','Rate'])
print('등분산 검정통계량=%.2f 유의확률=%.3f' %(ts,pvalue))
```

```
등분산 검정통계량=0.67 유의확률=0.413
```

통계적 가설검정

- 귀무가설 : 남여 CEO 대출 이자율은 동일하다. $H_0 : \mu_x = \mu_y$
- 대립가설 : 여성 대출이자율은 남성보가 높다. $H_1 : \mu_x > \mu_y$
- 검정통계량 : [등분산 가정 만족시] $\dfrac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{s_p^2(\frac{1}{n}+\frac{1}{m})}} \sim t(N+m-2)$

```
sp=((n-1)*std_x**2+(m-1)*std_y**2)/(n+m-2)
ts=(mean_x-mean_y)/np.sqrt(sp/n+sp/m)
pvalue=1-stat.t.cdf(np.abs(ts),n+m-2)
print('검정통계량=%.2f, 유의확률=%.6f)' %(ts,pvalue))
```

```
검정통계량=3.89, 유의확률=0.000052)
```

함수이용 : 등분산 가정이 만족하지 않으면 equal_var=False 옵션을 사용하면 된다.

```
import scipy.stats as stat
ts,p_value=stat.ttest_ind(df.loc[df['Gender']=='Female','Rate'],d
f.loc[df['Gender']=='Male','Rate'],equal_var=True)
print('검정통계량=%.2f, 유의확률=%.6f)' %(ts,p_value/2))
```

```
검정통계량=3.89, 유의확률=0.000052)
```

결론

유의확률 0.00005로 매우 작으므로 귀무가설은 기각되어 여성 CEO 대출 이자율이 남성 CEO보다 높으므로 은행은 대출 이자율 측면에서 여성 CEO를 차별했다고 할 수 있다.

CEO	평균	표준편차
여성	1.55%	0.637%
남성	1.28%	0.665%

연습문제 http://wolfpack.hnu.ac.kr/Stat_Notes/elem_stat/Stat_methods/newspaper.csv

도시 환경과에서 신문 수거 업체를 활용하고자 한다. 수거 업체는 가구당 하루 2파운드 이상 배출해야 수 거하여 수익이 발생한다고 하였다. 환경과에서 수거 업체의 수익을 보장할 수 있는지 알아보기 위하여 148가구에 대하여 조사한 자료이다.

도시(city), 근교(suburb) 지역간 신문수거 양의 차이가 있는지 1)상자그림과 히스토그램을 그리고(이상치가 있는 경우 제외) 2)개별 지역 95% 신뢰구간을 구하고 3)유의수준 5%에서 가설 검정하시오.

연습문제 : Major League Baseball Data from the 1986 and 1987 seasons.

https://vincentarelbundock.github.io/Rdatasets/csv/ISLR/Hitters.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 322 entries, 0 to 321
Data columns (total 21 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   322 non-null    object
 1   AtBat        322 non-null    int64
 2   Hits         322 non-null    int64
 3   HmRun        322 non-null    int64
 4   Runs         322 non-null    int64
 5   RBI          322 non-null    int64
 6   Walks        322 non-null    int64
 7   Years        322 non-null    int64
 8   CAtBat       322 non-null    int64
 9   CHits        322 non-null    int64
 10  CHmRun       322 non-null    int64
 11  CRuns        322 non-null    int64
 12  CRBI         322 non-null    int64
 13  CWalks       322 non-null    int64
 14  League       322 non-null    object
 15  Division     322 non-null    object
 16  PutOuts      322 non-null    int64
 17  Assists      322 non-null    int64
 18  Errors       322 non-null    int64
 19  Salary       263 non-null    float64
 20  NewLeague    322 non-null    object
dtypes: float64(1), int64(16), object(4)
```

1. 1987년 연봉 상위 20명, 그 다음 상위 20명(21번~40번)을 선택하여 고액연봉 집단과 세컨티어(second tier)집단으로 나누시오.
2. 타율(batting average)을 계산하시오. BA=Hits/AtBat
3. 집단에 대한 타율 히스토그램과 상자그림을 동시에 그리고 이상치를 진단하시오. 이상치가 존재하면 제외하고 향후 분석을 시작하시오.
4. 각 개별집단의 타율 평균에 대한 95% 신뢰구간을 구하시오.
5. 고액연봉 집단의 타율이 세컨티어 집단보다 높은지 유의수준 5%에서 검정하시오.

모분산 추론

1개 모집단

설정

확률표본은 모집단의 확률분포함수가 정규분포를 따른다. $(x_1, x_2, \ldots, x_n) \sim N(\mu, \sigma^2)$

정규분포를 따르지 않으면 통계량, MVUE 샘플링분포가 카이제곱 분포를 따르지 않는다.

모수 $\theta = \sigma^2$에 대한 MVUE 추정량

점 추정값 : $\hat{\sigma^2} = s^2$ (불편 추정량)=> $E(s^2) = \sigma^2$

추정분산 식은 매우 복잡하고 통계량 표본분산의 샘플링분포는 표준오차를 가지고 있지 않음

$\hat{\theta} = s^2$ 샘플링분포

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

가설검정 $H_0 : \sigma^2 = \sigma_0^2$

$$ts = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1)$$

예제

은행 문제에서 타 은행의 대출 이자율 분산은 0.5%이다. 동일한지 검정하시오. 단 이상치 제외

```
import scipy.stats as st
df=bank[bank['Rate']<3.29]
sigma0=0.5
n=df.shape[0]
s2=df['Rate'].var()
ts=(n-1)*s2/sigma0
p_value=st.chi2.cdf(ts,n-1)
print('표본분산=%.3f, 검정통계량=%.2f, 유의확률=%.6f)' %(s2,ts,p_value))
```

```
표본분산=0.444, 검정통계량=933.88, 유의확률=0.004127)
```

귀무가설이 기각되어 타 은행 분산 0.5% 비해 이 은행의 대출 이자율 분산은 0.444%로 유의적으로 낮다.

독립인 두 모집단

설정

확률표본은 모집단의 확률분포함수가 정규분포를 따른다. $(x_1, x_2, \ldots, x_n) \sim N(\mu_x, \sigma_x^2)$

독립인 다른 확률표본도 모집단의 확률분포함수가 정규분포를 따른다. $(y_1, y_2, \ldots, y_m) \sim N(\mu_y, \sigma_y^2)$

모수 $\theta = \frac{\sigma_x^2}{\sigma_y^2}$에 대한 MVUE 추정량

점 추정값 : $\hat{\theta} = \frac{s_x^2}{s_y^2}$: 불편추정량 => $E(\hat{\theta}) = \frac{\sigma_x^2}{\sigma_y^2}$

추정분산 식은 매우 복잡하고 통계량 표본분산 비의 샘플링분포는 표준오차를 가지고 있지 않음

$\hat{\theta} = \frac{s_x^2}{s_y^2}$ 샘플링분포 & 가설검정 $H_0 : \sigma_x^2 = \sigma_y^2$

$$ts = \frac{s_x^2}{s_y^2} \sim F(n-1, m-1)$$

예제

은행 문제에서 남녀별 대출 이자율 분산의 차이가 있는지 유의수준 5%에서 검정하시오. 단 이상치는 제외

```
import scipy.stats as st
df=bank[bank['Rate']<3.29]
ts,pvalue=st.levene(df.loc[df['Gender']=='Male','Rate'],df.loc[df
['Gender']=='Female','Rate'])
print('등분산 검정통계량=%.2f 유의확률=%.3f' %(ts,pvalue))
```

```
등분산 검정통계량=0.67 유의확률=0.413
```

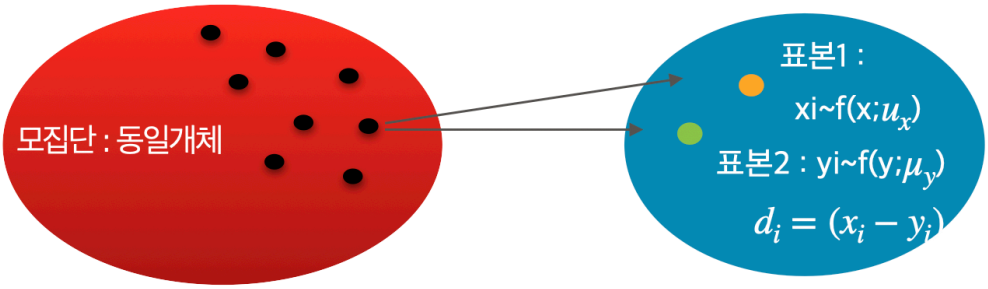
남성 은행 대출 이자율 분산은 0.405%, 여성은 0.442%로 귀무가설이 채택되어 등분산 가정(귀무가설 : 두 집단의 분산은 동일하다)이 만족된다.

```
df.groupby('Gender')['Rate'].var()
```

```
Female    0.405413
Male      0.441839
```

Snedecor, G. W., & Cochran, W. G. (1989). Statistical methods (8th ed.). Ames, IA: Iowa State University Press.

짝진집단 paired test



$$d_i \sim f(d;\mu_d)$$

1표본 데이터

$d_1 = -1.1, d_2 = 0.3, \ldots, d_n = -0.11$
표본크기 n인 1집단 평균 분석과 동일

개요

쌍으로 된 관측치의 차이를 구하면, $d_i = x_i - y_i$ 1집단 평균에 대한 추론과 동일하다.

모수 $\theta = \mu_d$ 및 점추정

$$\hat{\theta} = \sum d_i/n = \bar{D}$$

- 추정분산 : $V(\bar{D}) = \frac{\sigma_d^2}{n}$ => 표준오차 : $s(\bar{D}) = \frac{\sigma_d}{\sqrt{n}}$, 만약 σ_d 모르면 표준편차 s_d 로 대체한다.

샘플링 분포

$$\frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \sim N(0,1) = z$$

사례연구 : 'http://wolfpack.hnu.ac.kr/Stat_Notes/elem_stat/Stat_methods/mba2.csv')

MBA 전공 재무, 마케팅 연봉을 조사한 자료이다. 성적(GPA)에 따른 차이가 있을 가능성을 고려 하여 (4, 3.92)=1그룹, (3.92, 3,84)=2그룹2, ..., 총 25개 그룹에서 한 명씩 임의추출하여 연봉을 조사하였다. 유의수준 5%에서 재무전공, 마케팅 전공의 연봉 차이가 있는지 검정하시오. [Keller"ManagerialStatistics"9th edition]

사례연구 : http://203.247.53.31/2015_Fall/D4BE/ambulance.csv

소방차에 응급장비를 장착하는 문제에 대한 논쟁이 벌어져, (1) 콜센터 전화를 받고 소방차가 응급 차보다 1분 먼저 도착하고 (2) 8분 이내 도착하는 비율이 소방차가 높다면 응급장비를 소방차에 장 착하기로 시의회에서 결정하였다. (Cambridge, Waterloo, Kitchner) 3개 지역 중 장착 가능한 도시? [Keller"ManagerialStatistics"9th edition]