연구문제

> H시리얼 회사에서 작년 조사 결과 시리얼 구입 이유 - 다이어트 효과 때문 35%이었다. 그리고 아침 식사 비용 평균은 10.6$이었다. 올해 트랜드가 바뀌었는지 알아보기 위하여 다음 조사를 하였다. 아침으로 시리얼을 구입하는 고객 1,250명을 대상으로 시리얼 구입 할 때 고려하는 이유와 1)건강음식을 위하여 2)다이어트 효과 3)식이요법 4)가격고려, 그리고 시리얼 구입 비용을 조사하였다. http://203.247.53.31/2015_Fall/D4BE/cereal.csv

[데이터 불러오기]

```
1  import pandas as pd
2  df=pd.read_csv('http://203.247.53.31/2015_Fall/D4BE/cereal.csv')
3  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1250 entries, 0 to 1249
Data columns (total 2 columns):
Group    1250 non-null int64
Spend    1250 non-null float64
dtypes: float64(1), int64(1)
memory usage: 19.6 KB
```

[모비율 추론]

1. [신뢰구간] 다이어트 효과 이유 시리얼 구입 비율에 대한 95% 신뢰구간
2. [가설검정] 올해 다이어트 효과 이유 구입 비율은 작년(3%)과 동일한가?

모비율 추정

> normalize=True : 상대빈도, 디폴트=빈도 / sort=True : 빈도(상대빈도) 크기 순으로 정렬

```
1  df.Group.value_counts(normalize=True,sort=False) #normalize는 상대빈도=확률
```

```
1    0.2152
2    0.3872
3    0.1928
4    0.2048
Name: Group, dtype: float64
```

> [다이어트 이유 구입 비율] phat=38.7%

모비율(올해 다이어트 효과 이유 구입 비율) 95% 신뢰구간

phat=# of '관심범주'/(n=표본크기)

중심극한 정리 : paht ~ N(p,sqrt(p*(1-p))/n)

phat +- z(1-alpha/2)* sqrt(phat*(1-phat)/n)

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

> > normal : asymptotic normal approximation
>
> > agresti_coull : Agresti-Coull interval
>
> > beta : Clopper-Pearson interval based on Beta distribution
>
> > wilson : Wilson Score interval - p가 매우 작은 경우
>
> > jeffreys : Jeffreys Bayesian Interval
>
> > binom_test : experimental, inversion of binom_test - 소표본 min(np, nq)<=5

```
import statsmodels.stats.proportion as smp
lb,ub=smp.proportion_confint(484,1250,0.05)
print('phat=%.3f | (LB=%.3f, UB=%.4f)'%(484/1250,lb,ub))
```

```
phat=0.387 | (LB=0.360, UB=0.4142)
```

> > 0.35가 95% 신뢰구간에 포함되지 않으므로 작년 대비 다이어트 이유 구매 비율 증가(올해 38.7%)하였다.

가설검정

귀무가설 : p=p0 올해 다이어트 이유 구입 비율은 35%이다.

대립가설 : p^=p0 올해 다이어트 이유 구입 비율은 35%와 다르다.

검정통계량 TS=(표본비율phat-p0)/sqrt(p0*(1-p0)/n) ~ N(0,1) (z)

> statsmodels.stats.proportion.proportions_ztest(count, nobs, value=p0(귀무가설 p), alternative='two-sided', prop_var=False(a모집단 비율 아는 경우))

> > alternativestring in ['two-sided', 'smaller', 'larger']

```
import statsmodels.stats.proportion as smp
ts,p=smp.proportions_ztest(484,1250,0.35,alternative='two-sided')
print('phat=%.3f, 검정통계량=%.3f, 유의확률=%.4f'%(484/1250,ts,p))
```

```
phat=0.387, 검정통계량=2.700, 유의확률=0.0069
```

> > 귀무가설 기각 - 작년 대비 다이어트 효과를 위한 구입 비율이 38.7%로 유의한 증가하였음.

> > *** 양측검정에서 귀무가설이 기각되면, 단측 대립가설 검정에서도 귀무가설 기각된다.

> > *** 그러므로 양측대립가설 검정 결과 유의확률이 10% 이하이면, 단측 대립가설을 설정하는 경우에는 5%에서 귀무가설은 기각된다.

[모평균 추론]

올해 아침식사 평균 지출 비용에 대한 95% 신뢰구간

작년 평균 지출 비용 10.6$에 비해 지출의 변동이 있나? 가설 검정

[숫자 요약]

```
import pandas as pd
df=pd.read_csv('http://203.247.53.31/2015_Fall/P4BE/cereal.csv')
```

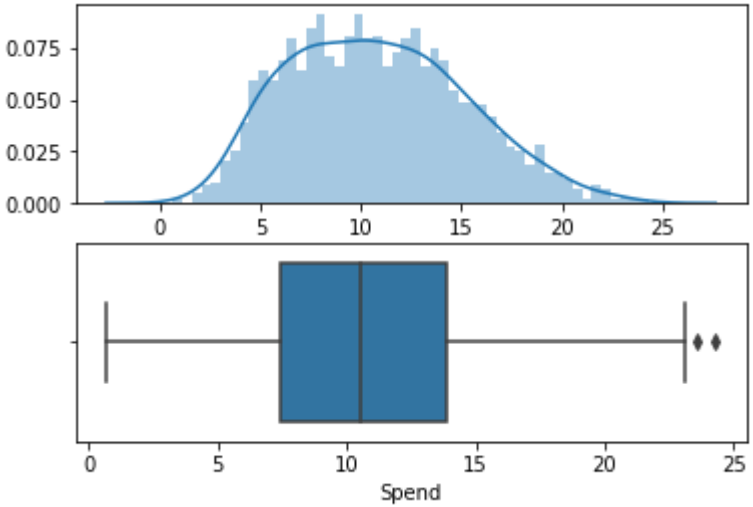
```
df=pd.read_csv( http://203.247.53.31/2015_Fall/D4BE/cereal.csv )
df.Spend.describe()
```

```
count    1250.000000
mean       10.807136
std         4.353315
min         0.670000
25%         7.432500
50%        10.520000
75%        13.867500
max        24.290000
Name: Spend, dtype: float64
```

[그래프요약]

```
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(1)
plt.subplot(211)
sns.distplot(df.Spend,bins=50)
plt.subplot(212)
sns.boxplot(df.Spend)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb55cd83048>
```



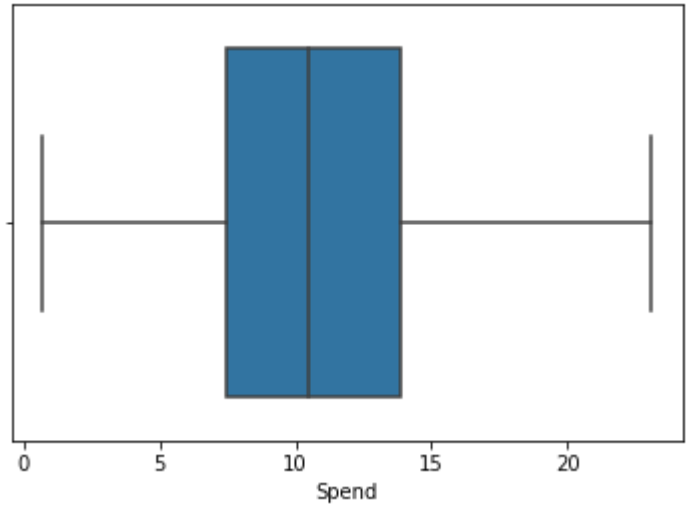
> 아침식사 비용 상한 이상치 2명 진단됨

이상치 제거

```
Q1=df.Spend.quantile(0.25)
Q3=df.Spend.quantile(0.75)
df_clean=df[(df.Spend>(Q1-1.5*(Q3-Q1)))&(df.Spend<(Q3+1.5*(Q3-Q1)))]
```

```
sns.boxplot(df_clean.Spend)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb55cb69630>
```



이상치 제거 후 기초통계량

```
df_clean.Spend.describe()
```

```
count    1248.000000
mean       10.786130
std         4.324987
min         0.670000
25%         7.427500
50%        10.505000
75%        13.837500
max        23.090000
Name: Spend, dtype: float64
```

> 올해 아침식사 비용 평균=10.79, 표준편차 = 4.325

[신뢰구간] 아침 식사 평균지출비용

```
import numpy as np
import scipy.stats
n=df_clean.shape[0];m=df_clean.Spend.mean();sd=df_clean.Spend.std()
lb=m-scipy.stats.t.ppf(0.975,n-1)*sd/np.sqrt(n)
ub=m+scipy.stats.t.ppf(0.975,n-1)*sd/np.sqrt(n)
print('표본평균=%.2f (하한=%.3f, 상한=%.3f)' %(m,lb,ub))
```

```
표본평균=10.79 (하한=10.546, 상한=11.026)
```

[가설검정]

> 귀무가설 : 올해 아침 식사 평균 지출비용은 10.6$이다. mu=10.6

> 대립가설 : 올해 아침 식사 평균 지출비용은 10.6$이 아니다. mu^=10.6

```
from scipy import stats
stats.ttest_1samp(df_clean.Spend,10.6)
```

```
Ttest_1sampResult(statistic=1.520331948112864, pvalue=0.12868107155635078)
```

귀무가설 채택. 올해 평균 비용 10.78$로 작년에 비해 조금 상승하였으나 유의적 상승은 아니다.

*** 대립가설을 작년에 비해 상승했다. mu>10.6 이라 하여도 유의확률은 0.129/2=0.065 여전히 5보다 크므로 귀무가설은 기각되지 못한다.