8.5 스튜던트 t분포, F카이제곱분포, F분포

이 섹션에서는 정규분포보다 꼬리가 두꺼운 스튜던트 t분포(Student's t distribution)와 카이제곱분포(chi-square distribution)와 F분포(F distribution)를 소개하고 이들 분포들이 어떻게 정규분포와 관련되는지 설명한다. 이들 분포들은 통계학에서 가설검정 등에 사용되는 중요한 분포들이다.

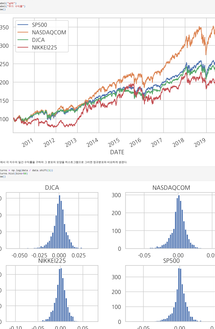
스튜던트 t분포

스튜던트 t분포는 표본평균 분포(sampling distribution)의 일종이다. 일반적으로 정규분포 분포의 꼬리 부분이 두꺼운 분포이다. 정규분포의 기대값을 알고 있지만 분산은 모를 때 표본평균의 기대값을 이용하여 모집단 분산을 추정하는 방법을 학습하게 된다. 이 때 표본평균 분포가 정규분포 분포와 비슷하지만 꼬리 부분이 두꺼운 분포를 가지게 된다. 이 때 이 분포를 스튜던트 t분포라고 한다. 이 때 표본평균 분포가 정규분포 분포와 비슷하지만 꼬리 부분이 두꺼운 분포를 가지게 되는 이유는 표본평균 분포의 분산을 추정하는 과정에서 발생하는 오차 때문이다. 이 때 표본평균 분포가 정규분포 분포와 비슷하지만 꼬리 부분이 두꺼운 분포를 가지게 되는 이유는 표본평균 분포의 분산을 추정하는 과정에서 발생하는 오차 때문이다. 이 때 표본평균 분포가 정규분포 분포와 비슷하지만 꼬리 부분이 두꺼운 분포를 가지게 되는 이유는 표본평균 분포의 분산을 추정하는 과정에서 발생하는 오차 때문이다.

실제로 분포 위의 확률이 정규분포(Normal), 스튜던트 t분포(t dist), 스튜던트t(Normal)와 스튜던트t(t dist)의 경우 차이가 있는지 비교하기 위해 다음 그림을 그려보자. 이 그림을 통해 차이가 있는지 확인할 수 있다.

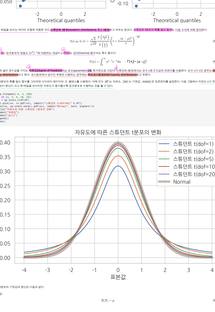
```
import pandas, pandas.plot, matplotlib, numpy, sys, os
%matplotlib inline
symbols = ['SP500', 'NASDAQCOM', 'DJCA', 'NIKKEI225']
start = '1990-01-01'
for sym in symbols:
    data = pandas.DataReader(sym, 'yahoo', start)
    exec('%s = data["Adj Close"]' % sym)
    print('%s: %d rows' % (sym, len(data)))
pd = pandas.DataFrame(
    {'SP500': SP500,
     'NASDAQCOM': NASDAQCOM,
     'DJCA': DJCA,
     'NIKKEI225': NIKKEI225})
pd.plot(figsize=(8, 5))
pd.ylabel('주가 수익률')
pd.xlabel('DATE')
```

```
SP500: 7328 rows
NASDAQCOM: 7328 rows
DJCA: 7328 rows
NIKKEI225: 7328 rows
```



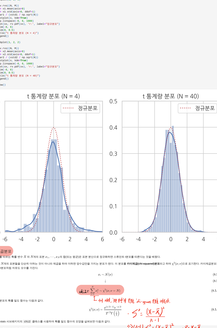
이 데이터에서 주가의 일일 수익률을 구하기 위 결과와 정규분포 히스토그램을 그리기 위해 정규분포의 시각화를 한다.

```
log = lambda x: numpy.log(x)
for sym in symbols:
    exec('%s = log(%s[1:] / %s[:-1])' % (sym,sym,sym))
pd = pandas.DataFrame(
    {'SP500': SP500,
     'NASDAQCOM': NASDAQCOM,
     'DJCA': DJCA,
     'NIKKEI225': NIKKEI225})
pd.hist(bins=50, sharex=True, sharey=True, grid=False)
pd.ylabel('주가 수익률')
pd.xlabel('DATE')
```



위 차트는 실제 정규분포 분포와 비슷하지만 정규분포보다 꼬리 부분이 두꺼운 분포이다. 정규분포 분포와 비슷하지만 정규분포 분포와 다른 점이다.

```
from scipy.stats.mstats import mquantiles
for sym in symbols:
    exec('%s = %s[%s.isnull() == False]' % (sym,sym,sym))
    exec('%s = %s[%s < mquantiles(%s, prob=[0.01])]' % (sym,sym,sym,sym))
    exec('%s = %s[%s > mquantiles(%s, prob=[0.99])]' % (sym,sym,sym,sym))
    exec('%s = %s - %s.mean()' % (sym,sym,sym))
pd = pandas.DataFrame(
    {'SP500': SP500,
     'NASDAQCOM': NASDAQCOM,
     'DJCA': DJCA,
     'NIKKEI225': NIKKEI225})
pd.hist(bins=50, sharex=True, sharey=True, grid=False)
pd.ylabel('주가 수익률')
pd.xlabel('DATE')
```



위 그림에서 확률분포 분포와 정규분포 분포의 차이가 있는지 확인하기 위해 정규분포 분포와 비교하는 그림을 그려보자. 이 그림을 통해 차이가 있는지 확인할 수 있다.

$$f(x) = \int_0^\infty e^{-t} t^{x-1} dt \quad \text{(8.5.1)}$$

감마함수의 일종이 스튜던트 t분포의 밀도함수이다. 스튜던트t분포의 밀도함수는 다음과 같다. 이 때 $\Gamma(\cdot)$ 는 감마함수이다. 이 때 ν 는 스튜던트 t분포의 자유도를 의미한다. 자유도는 일반적으로 표본의 크기를 의미한다.

스튜던트 t분포의 확률 밀도 함수는 다음과 같다. 이 때 $\Gamma(\cdot)$ 는 감마함수이다. 이 때 ν 는 자유도이다. 일반적으로 자유도는 표본의 크기를 의미한다. 표본의 크기가 클 때 자유도가 커진다.

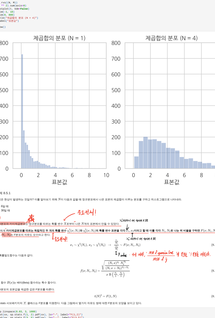
$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{(8.5.2)}$$

이 때 $\Gamma(\cdot)$ 는 감마함수이다. 표본 크기가 커질 때 스튜던트 t분포의 밀도함수는 다음과 같이 정규분포 밀도함수에 수렴한다. 이 때 ν 는 자유도이다.

$$\lim_{\nu\to\infty} f(x;\nu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{(8.5.3)}$$

이 때 ν 는 자유도이다. 표본 크기가 커질 때 스튜던트 t분포의 밀도함수는 다음과 같이 정규분포 밀도함수에 수렴한다.

```
x = numpy.linspace(-4, 4, 200)
for df in [1, 2, 5, 20]:
    exec('%s = %s.pdf(x, %d)' % ('t'+str(df), 'stats.t', df))
n = stats.norm(0, 1).pdf(x)
pd = pandas.DataFrame(
    {'t1': t1,
     't2': t2,
     't5': t5,
     't20': t20,
     'Normal': n})
pd.plot(figsize=(8, 5))
pd.ylabel('p(x)')
pd.xlabel('표본값')
pd.title('자유도에 따른 스튜던트 t분포의 변화')
```



스튜던트 t분포는 다음과 같은 성질을 가진다.

- 대칭성

$$f(x) = f(-x) \quad \text{(8.5.4)}$$

- 분산

$$\mathrm{Var}[X] = \frac{\nu}{\nu-2} \quad \text{(8.5.5)}$$

분산의 식에 $\nu - 2$ 인 것을 보아서 자유도는 $\nu > 2$ 일 때만 의미가 있다.

중요 개념

자유도가 클 때 스튜던트 t분포는 정규분포 분포와 비슷하다.